软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(10):2572−2585 [doi: 10.3724/SP.J.1001.2012.04181]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于 Deep Belief Nets 的中文名实体关系抽取*

陈 宇, 郑德权+, 赵铁军

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Chinese Relation Extraction Based on Deep Belief Nets

CHEN Yu, ZHENG De-Quan+, ZHAO Tie-Jun

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: dqzheng@mtlab.hit.edu.cn, http://www.hit.edu.cn

Chen Y, Zheng DQ, Zhao TJ. Chinese relation extraction based on Deep Belief Nets. *Journal of Software*, 2012,23(10):2572−2585 (in Chinese). http://www.jos.org.cn/1000-9825/4181.htm

Abstract: Relation extraction is a fundamental task in information extraction, which is to identify the semantic relationships between two entities in the text. In this paper, deep belief nets (DBN), which is a classifier of a combination of several unsupervised learning networks, named RBM (restricted Boltzmann machine) and a supervised learning network named BP (back-propagation), is presented to detect and classify the relationships among Chinese name entities. The RBM layers maintain as much information as possible when feature vectors are transferred to next layer. The BP layer is trained to classify the features generated by the last RBM layer. The experiments are conducted on the Automatic Content Extraction 2004 dataset. This paper proves that a character-based feature is more suitable for Chinese relation extraction than a word-based feature. In addition, the paper also performs a set of experiments to assess the Chinese relation extraction on different assumptions of an entity categorization feature. These experiments showed the comparison among models with correct entity types and imperfect entity type classified by DBN and without entity type. The results show that DBN is a successful approach in the high-dimensional-feature-space information extraction task. It outperforms state-of-the-art learning models such as SVM and back-propagation networks.

Key words: DBN (deep belief nets); neural network; relation extraction; deep architecture network; character-based feature

摘 要: 关系抽取是信息抽取的一项子任务,用以识别文本中实体之间的语义关系.提出一种利用 DBN(deep belief nets)模型进行基于特征的实体关系抽取方法,该模型是由多层无监督的 RBM(restricted Boltzmann machine)网络和一层有监督的 BP(back-propagation)网络组成的神经网络分类器. RBM 网络以确保特征向量映射达到最优,最后一层 BP 网络分类 RBM 网络的输出特征向量,从而训练实体关系分类器.在 ACE04 语料上进行的相关测试,一方面证明了字特征比词特征更适用于中文关系抽取任务;另一方面设计了 3 组不同的实验,分别使用正确的实体类别信息、通过实体类型分类器得到实体类型信息和不使用实体类型信息,用以比较实体类型信息对关系抽取效果的影响.实验结果表明,DBN 非常适用于基于高维空间特征的信息抽取任务,获得的效果比 SVM 和反向传播网络更好.

关键词: DBN(deep belief nets);神经网络;关系抽取;深层网络;字特征

* 基金项目: 国家自然科学基金(61073130); 国家高技术研究发展计划(863)(2011AA01A207)
收稿时间: 2011-06-16; 修改时间: 2011-08-09; 定稿时间: 2012-01-16

中图法分类号: TP391 文献标识码: A

信息抽取是指从大规模的无结构文本中提取出用户感兴趣的信息,并以结构化或者半结构化的形式输出,其主要包括 3 项子任务:实体抽取、关系抽取和事件抽取,本文主要关注第 2 项子任务.根据 Automatic Content Extraction(ACE)的定义,实体关系是实体之间显式或者隐式的语义联系.需要预先定义实体关系的类型,然后识别实体之间是否存在语义关系或是属于哪一种预定义的关系类型.例如,在“英国外长库克”这个描述中,“英国”和“库克”分别是两个实体,组成一个关系实例,它们之间的关系类型属于角色(role)关系.目前,关系抽取主要有基于模式匹配、基于特征和基于核函数这 3 种方法.

基于模式匹配的方法是根据已有的关系实例,人工构造出基于词法分析和句法分析等特征的模式集合;然后,将待识别的关系实例进行相同的预处理,并与模式集合中的模式进行匹配,如果匹配成功,即认为该实例具有对应模式的关系属性.基于模式匹配的方法能够取得较高的准确率,但是由于难以建立完整而准确的模式集合,很难取得理想的召回率.模式集合对语料的依赖性很强,如果语料的领域发生改变,模式需要大量改写甚至重写.模式扩展的方法可在一定程度上缓解基于模式匹配方法的不足,其思想是定义少量的关系模式种子集合,通过自学习,种子集合得到扩展[1].

基于特征的方法是将关系实例通过一定粒度的词法分析和句法分析转换为平面特征向量,然后采用 Maximum Entropy(ME)[2],Support Vector Machine(SVM)[3]等机器学习模型比较特征向量之间的相似性并分类.相对于建立模式集合,特征的提取简单、有效,不需要具有专业知识的专家进行大量人工操作.基于特征的信息抽取方法重点在于构造完整的特征和选取合适的机器学习模型,特征应最大限度地包含实例的信息,以提高特征的区分度.目前,特征的构造已经利用了大多数的自然语言处理方法,常用的特征组合包括词语特征、词性特征、实体属性特征、句法组合、语义特征以及结构化信息.词法与实体特征的提取相对简单,但是句法、语义以及结构化特征的提取受限于对原始语料进行句法分析、短语块标注等预处理工作的性能[4].另外,基于特征方法的缺点还在于未能引入语法结构和依赖信息.

基于核函数的方法是通过构造核函数,隐式地计算特征向量内积,从而得到关系实例之间的相似性.核函数引入的实例结构和依赖信息,对关系信息有重要的指示作用.近年来,多种不同的核函数运用在英文关系抽取任务中,已经取得优于基于特征的方法的结果,例如语义序列核函数[5]、依赖树核[6]、最短路径依赖核[7]和卷积语法树核[8].但是,由于中文的句子结构相对英语而言较为松散,词语之间没有位置指示信息,甚至中文的语法分析方法与工具的运用还不够成熟,所以基于核函数的方法在中文关系抽取任务中未能取得期望的效果.

关系抽取的研究大多集中于英文语料,对中文语料的研究相对较少,而且其难度要远大于英文.主要原因是[9]:① 汉语词语之间没有明显的分割标志;② 汉语中的词存在更多歧义现象;③ 汉语词语由字组合而成,组合的复杂度高;④ 汉语的词法语态信息没有英语丰富,例如,汉语词语没有时态、字母大小写的特征等;⑤ 目前,中文分词系统结果依然存在误差,分词错误将噪音引入关系抽取任务.Jing[10]提出利用基于字特征表征名词信息,克服了以上大部分的汉语难点.基于字特征使得不再需要对语料进行分词预处理,将字与字之间如何组成词交由机器学习模型去决定.我们针对 ACE04 的语料,将实体的字特征、实体的指称特征、实体的类别特征和实体之间的相对位置特征组成了关系实例的组合特征向量,利用 Deep Belief Net(DBN)神经网络进行关系抽取.

目前,绝大部分关系抽取的相关研究中都是假设关系中的实体已被正确识别,将正确的实体属性作为组合特征之一.由于识别实体的类型也是信息抽取的一项子任务,是关系抽取的基础,对于大多数文本语料,实体属性是未被标注的.要对未标注的语料进行关系抽取,必须先对语料进行实体识别,实体识别的结果直接影响关系抽取的结果.Claudio[11]分别假设实体的边界信息和类型信息未知,利用条件随机场(conditional random field,简称 CRF)模型先后识别实体的边界信息和类别信息,然后将这些信息加入核函数中进行关系抽取,验证了英语语料中实体抽取对关系抽取的重要性.实体类型信息对于实体关系具有重要的指示,例如,假设两个人名(person)实体之间存在关系,那么这个关系只能是 social 类型,不会存在于其他 ACE 预定义的关系类型中.本文设计了 3

组实验,假设实体的边界已知,分别利用准确的实体类型特征、有噪音的实体类型特征和不使用实体类型特征,让我们能够直观地理解在中文语料中实体类型信息对于关系抽取的影响.

本文提出利用 DBN 模型进行基于特征的实体关系抽取任务.DBN 模型结合了无监督学习和有监督学习的优点,是一种对高维稀疏特征向量具有强大分类能力的神经网络,它由若干层无监督的 Restricted Boltzmann Machine(RBM)网络和一层有监督的反向传播网络(back-propagation,简称 BP)组成.DBN 模型的训练过程分为两个阶段:首先利用多层 RBM 对特征集合进行聚类,然后利用 BP 对聚类结果进行分类,并同时对 RBM 网络进行微调.本文将 DBN 方法与 SVM 和传统神经网络方法(neural network,简称 NN)进行了比较,实验结果表明,DBN 优于这两种传统的机器学习方法,验证了 DBN 对关系抽取任务的有效性.

1 关系抽取

1.1 任务定义

事实存在的物体或者物体集合,我们称其为实体.实体是组成关系的基本元素.ACE 定义了 5 种实体类别,分别是:人(person)、组织机构(organization)、行政区(geo-political entity)、地点(location)、设施(facility).文本中对物体的语言描述称为指称,每种实体的指称有 3 种类型:命名性指称(name mention)、名词性指称(nominal mention)、代词性指称(pronoun mention).例如,以下是同一个人的 3 个指称:

- 命名性指称:比尔盖茨;
- 名词性指称:微软 CEO;
- 代词性指称:他.

每个中文句子可能包含有若干个实体,由于在不同句子中的两个实体基本不存在语义关系,本文只考虑同一句子中的两个实体之间存在的语义联系.形式化地表示为:一个句子 S 中含有实体集合$\{E_1,E_2,\ldots,E_n\}$,n 为实体在句子中出现的序号.关系抽取就是识别 E_i 和 E_j 之间的关系 R_{ij},如公式(1)所示:

$$(E_i,E_j,S)\rightarrow R_{ij} \tag{1}$$

我们称三元组(E_i,E_j,R_{ij})为一个关系实例.根据 ACE04 语料的定义,预先定义了 5 种实体关系:

- Role:人与组织机构、设施、或者行政区之间的角色从属关系;
- Part:组织机构、设施和行政区之间的整体与部分的关系;
- At:人、组织机构、行政区和设施与地点之间的位置关系;
- Near:人、组织机构、行政区和设施邻近一个地点或者行政区;
- Social:人与人之间的个人关系或者从属关系.

其中,这些关系可以分为方向相关和方向无关两种.对于方向相关的实体关系,$R_{ij}\neq R_{ji}$ 成立,即假设(E_i,E_j,R_{ij})成立,那么$(E_j,E_i,Null)$成立,$Null$ 表示实体间不存在预定义的语义关系.例如,“盖茨”是“微软”的角色(role)从属,但是“微软”不是“盖茨”的角色从属.对于方向不相关的实体关系,$R_{ij}=R_{ji}$ 成立,即(E_i,E_j,R_{ij})和(E_j,E_i,R_{ji})都成立.由 ACE 的定义可知,Role,Part,At 为方向相关的实体关系,Near 和 Social 为方向无关的实体关系.本文实验中并没有将实体的顺序作为组合特征之一,而是在实体关系被识别之后,根据实体属性,实体出现在句子中的相对位置对实体进行顺序修正的后处理.

1.2 组合特征选择

本文主要利用了实体的字特征、实体类型特征、实体指称特征和实体之间的相对位置特征.这些特征都易于提取,并且不引入噪音,准确无误,不受词法分析等预处理性能的影响.为了证明字特征比传统的词特征更适合中文关系抽取任务,我们用词特征和词性特征替换字特征,并将它们的结果进行比较.

(1) 字特征.字特征将实体看成是由若干单个字符组成,字是组成实体的基本单位,字与字如何组合成词语,由机器学习模型决定.例如:“老”与“李”可以组合成“老李”,并被分类为“人名”;“老”与“挝”可以组合成“老挝”,并被分类为“行政区”.在高维的特征向量中,这种组合方式是极其复杂的.本文将实体中

出现的字组成词典 $D=\{d_1,d_2,\ldots,d_n\}$,d_i 表示字,$i\in[1,n]$,词典 D 包括了所有出现在关系实例中的字.将每一个实体 E 的基于字的特征向量表示为 $V(E)=\{v_1,v_2,\ldots,v_n\}$.特征向量与词典具有相同的维数,其中,$v_i$ 的值满足等式(2):

$$v_i=\begin{cases}1, & d_i\in E\\ 0, & d_i\notin E\end{cases} \tag{2}$$

(2) 词特征和词性特征.词特征是将词作为组成实体的基本单位,其提取过程与字特征相似,其词典是由词组成.本文利用中国科学院的开源分词系统 ICTCLS 对实体进行分词预处理和词性标注;

(3) 实体类型特征.ACE04 语料中,预定义的实体类型有 5 类,每一个实体属于并且只属于其中一类;

(4) 实体指称特征.ACE04 语料中,预定义的实体指称有 3 类,每一个实体属于并且只属于其中一类;

(5) 实体相对位置特征.Zhang[12]证明了实体相对位置特征是关系抽取重要的组合特征.我们定义了 3 种实体间的相对位置类型,分别是嵌套、相邻和分离.嵌套表示其中一个实体嵌套在另一个实体之中;相邻表示两个实体连接在一起,它们之间没有其他字符;分离表示两个实体之间存在其他字符.假设 $E_i.start$ 和 $E_i.end$ 分别表示实体 E_i 的开始位置和结束位置,表 1 列出了关系实例中两个实体相对位置的形式化定义.

Table 1 Internal position structure features between two named entities

表 1 实体的相对位置

类型	成立条件
嵌套	$(E_i.start,E_i.end)\supset(E_j.start,E_j.end)$
相邻	$E_i.end=E_j.start-1$
分离	$(E_i.start<E_j.start)\&(E_i.end+1<E_j.start)$

1.3 本文的工作

本文的主要任务有 3 个,实现过程如图 1 所示.(1) 验证 DBN 是一个适合于信息抽取的机器学习模型;(2) 比较基于字特征和基于词特征的优劣性;(3) 检验实体的类别信息对关系抽取的影响.

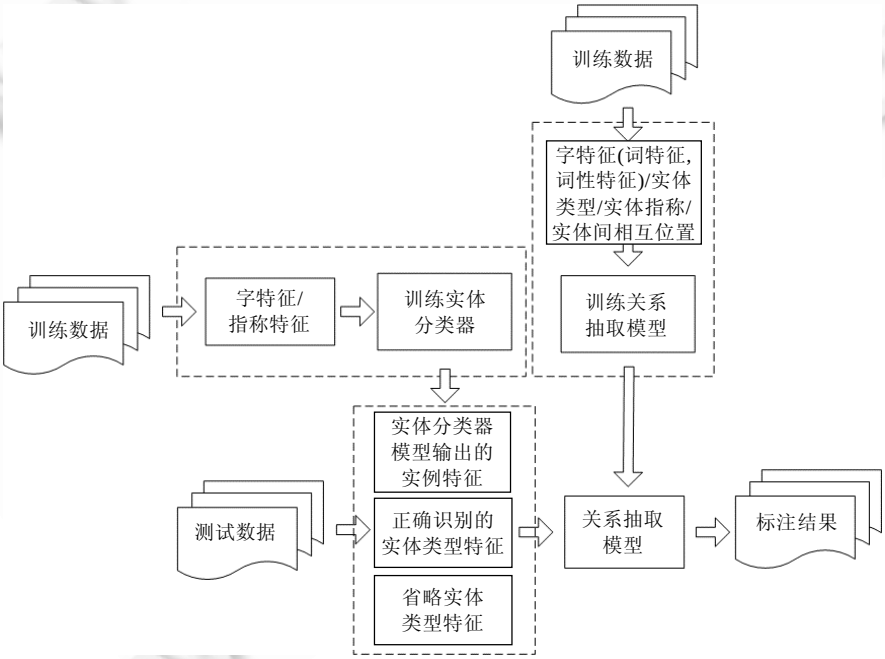


Fig.1 System's architecture

图 1 系统的架构

为了实现任务(1),我们假设实体的边界与类型信息已知,将字特征、实体类型特征、实体指称特征和实体间的相对位置关系作为关系实例的组合特征.将训练数据的特征集合分别用以训练 DBN,SVM 和 NN 的实体关系分类器,然后用训练完毕的模型对测试数据进行测试.这样就可以比较在相同特征组合的数据集合下,各个不同模型实体关系分类器的效果.

对于任务(2),中文的词是表达语义的基本单位,由若干个字组成,并且词与词之间没有分割标志.信息抽取任务的传统做法是:首先对语料进行分词的预处理,得到词语和词语的词性.但是中文分词系统依然存在一定的误差,使用词特征或词性特征会将噪音引入关系抽取系统.为了验证字特征和词特征哪一个更适合于关系抽取任务,我们将上一组实验的字特征替换成词特征和词性特征的组合,并将它们的结果进行比较.

对于任务(3),本文设计了 3 组实验验证实体的类型信息对关系抽取的影响:第 1 组实验假设关系中的实体的类型已被正确识别,将其作为关系实例特征向量的组合特征之一;第 2 组实验是先利用 DBN,SVM 和 NN 模型建立实体类型分类器,利用此分类器先得到关系实例中的实体的类型,然后将这种有噪音的实体类型特征作为关系实例特征向量的组合元素之一.训练实体类型分类器的过程与训练实体关系分类器的过程相似,但是后者的输入数据只是由字特征和实体指称特征组成;第 3 组实验是摒弃实体的类型信息,不将其加入关系实例的组合特征中.

2 研究模型

2.1 DBN神经网络

DBN 是由若干层 RBM 和一层 BP 组成的一种深层神经网络[13],其结构如图 2 所示.DBN 在训练模型的过程中主要分为两步:第 1 步,分别单独无监督地训练每一层 RBM 网络,确保特征向量映射到不同特征空间时,都尽可能多地保留特征信息;第 2 步是在 DBN 的最后一层设置 BP 网络,接收 RBM 的输出特征向量作为它的输入特征向量,有监督地训练实体关系分类器.而且每一层 RBM 网络只能确保自身层内的权值对该层特征向量映射达到最优,并不是对整个 DBN 的特征向量映射达到最优,所以反向传播网络还将错误信息自顶向下传播至每一层 RBM,微调整个 DBN 网络.RBM 网络训练模型的过程可以看作对一个深层 BP 网络权值参数的初始化,使 DBN 克服了 BP 网络因随机初始化权值参数而容易陷入局部最优和训练时间长的缺点.

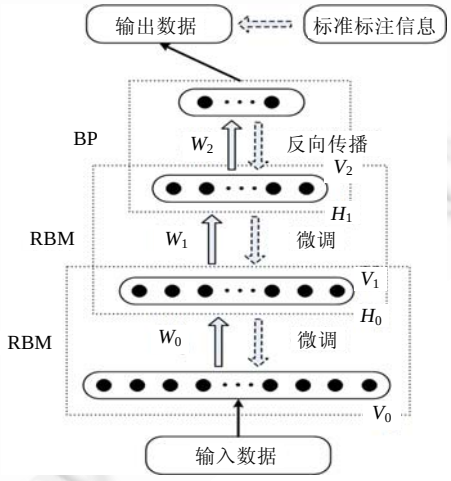


Fig.2 Structure of a DBN

图 2 DBN 网络的结构图

DBN 是一种深层神经网络,底层的神经网络接收原始的特征向量,在自底向上的传递过程中,从具体的特征向量逐渐转化为抽象的特征向量,在顶层的神经网络形成更易于分类的组合特征向量.增加网络层数能够将特

征向量更加抽象化[14].而且,虽然 RBM 确保训练后的层内参数达到最优,但却不能完全消除映射过程中产生的错误和不重要的信息,多层神经网络的每一层网络会弱化上一层网络产生的错误信息和次要信息,因此,深层网络较单层网络精确度更高.

2.2 RBM神经网络

RBM 是 DBN 的核心组件之一,它由一个可见层 V 和一个隐含层 H 组成,层间的节点两两相连,层内的节点不相连,其结构如图 3 所示.

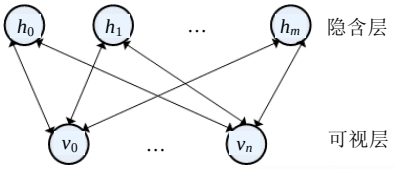


Fig.3 Structure of a RBM

图 3 RBM 网络的结构图

令 v_i 和 h_j 分别表示可视层和隐含层内的节点值,b 和 c 分别表示可视层和隐含层的偏置量,W 表示可视层和隐含层之间的权值.利用公式(3)可以由已知的可视层的节点值得到隐含层的节点值:

$$p(h_j=1)=\frac{1}{1+\exp(-b_j-\sum_i v_i w_{ij})} \tag{3}$$

RBM 是对称网络,同理,利用公式(4)可以由已知的隐含层的节点值得到可视层的节点值:

$$p(v_i=1)=\frac{1}{1+\exp(-c_i-\sum_j h_j w_{ji})} \tag{4}$$

那么,可视层内的特征向量 v 和隐含层内的特征向量 h 的联合概率分布满足:

$$p(v,h)\propto \exp(-E(v,h))=\mathrm{e}^{h^T W v+b^T v+c^T h} \tag{5}$$

其中,$E(v,h)$是特征向量 v 和特征向量 h 数学期望,其绝对值的大小代表特征向量 h 保存着特征向量 v 的信息的多少,需要确定的参数为$\theta=(W,b,c)$,其中,W 是 RBM 的权值参数,b 是可视层的偏置量,c 是隐含层的偏置量,使得联合概率分布 $P(v,h)$最大[15].最大似然法并不能求出满足条件的参数θ,传统的做法是利用马尔可夫链蒙特卡罗(Markov chain Monte Carlo,简称 MCMC).MCMC 的特性使得可视层和隐含层互为条件,不断地求得更新状态,最后它们共同趋向平稳状态,而此时的 $P(v,h)$达到最大[16].此后可以求得最大联合概率分布与初始状态的联合概率分布的斜率 $\frac{\partial \log P(v,h)}{\partial \theta}$,然后用公式(6)更新权值$\theta$.

$$\theta^{(\tau+1)}=\theta^{(\tau)}+\eta\frac{\partial \log P(v,h)}{\partial \theta}\bigg|_{\theta^{\tau}} \tag{6}$$

其中,τ为迭代次数,η为学习速度.其过程如图 4 所示.

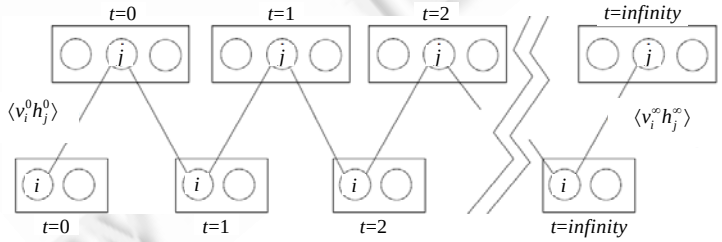


Fig.4 Process of a RBM using Monte Carlo Markov chain

图 4 基于马尔可夫链的 RBM 网络自训练过程

v^0 是 t=0 时刻可视层的特征向量,即是 RBM 的输入向量;h^0 是由 v^0 根据公式(3)得到的隐含层特征向量;v^1 是 t=1 时刻可视层的特征向量,根据 h^0 由公式(4)计算得到.以此类推,v^∞和 h^∞分别是 t=∞时刻可视层和隐含层的特征向量.斜率可由公式(7)计算得出:

$$\frac{\partial \log p(v,h)}{\partial \theta_{ij}} = \langle h_j^0(v_i^0 - v_i^1)\rangle + \langle v_i^1(h_j^0 - h_j^1)\rangle + ... = \langle h_j^0 v_i^0\rangle - \langle h_j^0 v_i^1\rangle + \langle v_i^1 h_j^0\rangle - \langle v_i^1 h_j^1\rangle + ... = \langle h_j^0 v_i^0\rangle - \langle h_j^\infty v_i^\infty\rangle \quad (7)$$

其中,$\langle h^0 v^0\rangle$为输入特征向量与其对应的隐含层特征向量的点乘的平均值;$\langle h^\infty v^\infty\rangle$为马尔可夫链末端可视层特征向量与其对应的隐含层特征向量的乘积的平均值,$\langle h^\infty v^\infty\rangle$是收敛的.由公式(7)可知,联合概率分布的斜率与中间状态无关,只与网络的初始状态和最终状态有关.根据公式(6)可以得出修改后的参数 θ,从而达到自训练的目的.

2.3 Contrastive divergence准则

利用马尔可夫链的方法求最佳联合概率 $P(v^\infty,h^\infty)$与初始联合概率分布 $P(v,h)$,收敛速度很难保证,难以确定步长∞.Hinton[17]提出利用 Contrastive Divergence(CD)准则快速提高计算速度并且保持精度.利用 Kullback-Leibler 距离衡量两个概率分布的"差异性",表示为 $KL(P||P')$,如公式(8)所示.

$$CD_n = KL(p_0||p_\infty) - KL(p_n||p_\infty) \quad (8)$$

其中,P_0 为 RBM 网络初始状态的联合概率分布,P_n 为经过 n 步马尔可夫链之后的 RBM 网络的联合概率分布,P_∞ 为马尔可夫链末端的 RBM 网络的联合概率分布.所以,CD_n 可以看作是 P_n 衡量介于 P_0 和 P_∞之间的位置.不断地将 P_n 赋值给 P_0,得到新的 P_0 和 P_n.实验证明,在 r 次求斜率修正参数 θ后,CD_n 必将趋向于 0,且精度近似于马尔可夫链.本文实验中[18],设定 n=1,图 5 给出了 RBM 的训练过程,其实现步骤如算法 1 所示.

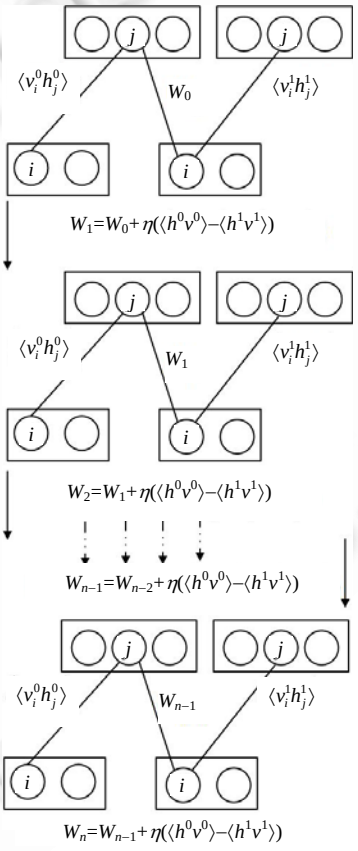


Fig.5 Process of a RBM using contrastive divergence

图 5 基于 CD 准则的 RBM 网络自训练过程

算法 1. 基于 CD 准则的 RBM 网络自训练过程.

(1) 随机初始化$\theta_0=(W_0,b_0,c_0)$赋值给θ_t,并设定迭代次数 *Step*;

(2) 将输入特征向量赋值给 v^0,并利用公式(3)和公式(4)计算特征向量 h^0,v^1 和 h^1;

(3) 利用公式(7)得到 RBM 网络初始状态与更新状态下的联合概率分布的斜率,并代入公式(7)修正参数θ_t,得到θ_{t+1};

(4) 如果 $t=Step$,程序结束;如果 $t<Step$,则将θ_{t+1}赋值于θ_t,并转步骤(2).

2.4 BP神经网络

BP 网络是一种有监督分类器,用于将 RBM 提取的特征向量进行分类,并起到微调整个 DBN 的作用.其训练过程主要分为两步:第 1 步是前向传播,将输入特征向量沿输入端传播至输出端;第 2 步是反向传播,将 BP 网络的输出结果与正确结果相比较得到误差,然后将误差从输出端反向传播至输入端,以修改 DBN 的参数.本文实验中利用 sigmod 函数作为 BP 的网络节点的求值函数,其实现步骤如算法 2 所示.

算法 2. BP 网络的训练过程.

(1) 随机初始化顶层反向传播网络的参数,设定训练步长为 *N*;

(2) 进行前向计算,对第 l 层的 j 单元节点,其值为 $y_j^l(n)=\sum w_{ji}(n)y_i^{l-1}(n)$, 若神经元 j 属于输出层($l=L$),则令 $y_j^L(n)=O_j(n)$,误差 $e_j(n)=d_j(n)-O_j(n)$,d_j 为正确信息;

(3) 计算δ,将δ反向传递用以自顶向下修正网络的权值参数,对于输出单元: $\delta_j^l(n)=e_j(n)O_j(n)[1-O_j(n)]$;对于隐含层单元: $\delta_j^l(n)=y_j^l(n)[1-y_j^l(n)]\sum\delta_k^{l+1}(n)w_{kj}^{l+1}(n)$;

(4) 修改权值: $w_{ji}^l(n+1)=w_{ji}^l(n)+\eta\delta_j^l y_i^{l-1}(n)$,$\eta$为学习速率;

(5) 如果 $n=N$,则训练结束;反之,$n=n+1$,转步骤(2).

3 实验与分析

3.1 实验设置与评价

本文选用 ACE04 的语料作为实验数据,包含 221 篇消息文本,10 228 个实体,5 240 个关系实例.我们还从语料中提取了处在同一句子内却不包含关系的实体对 23 600 个.所有实体类型分类实验和关系抽取实验都采用 4 折交叉验证.我们将用准确率、召回率和 F 系数来评价关系抽取的结果.对于实体类型分类任务,由于假设实体与其边界已被正确识别,所以它的准确率、召回率和 F 系数是相等的.F 系数可由公式(9)求得.

$$F\text{系数}=\frac{2\times\text{准确率}\times\text{召回率}}{\text{准确率}+\text{召回率}} \tag{9}$$

为了验证本文提出方法的有效性,并与其他传统方法进行比较,本文设计了 3 项任务:

- 任务 1,验证本文提出的 DBN 方法比传统的 SVM 和 NN 方法更适合于信息抽取任务;
- 任务 2,验证基于字特征的方法优于传统的基于词特征的方法;
- 任务 3,检验实体的类型信息对关系抽取的影响.

针对任务 1,我们设计了 3 组不同的实验进行关系抽取:

- 实验 1,只识别关系实例,只识别候选实体关系对是否存在语义关系,不识别实体关系属于哪一种类型的关系;
- 实验 2,先识别关系实例再识别关系类型,对那些被识别为存在语义关系的关系实例进行分类,识别出它们分别属于哪一类;
- 实验 3,直接识别候选实体关系对的关系类型,除了语料中预定义的 5 种实体类型,将不存在关系的实体对定义为 *Null* 类型,认为它也是一类实体关系.将所有候选关系实例直接分类为这 6 类中的一类.

我们将字特征、实体指称特征、准确的实体类型特征和实体相对位置作为关系实例的组合特征,分别利用

DBN^i(i 表示 DBN 包含的 RBM 的层数),SVM 和 NN 这 3 种机器学习模型完成以上 3 组实验,并进行结果的比较.组合特征的维数为 3 015.

对于任务 2,本文将字特征替换为词特征和词性特征的组合.

- 首先,利用中国科学院的分词系统 ICTCLS (http://ictclas.org/)对候选关系实例中的实体进行分词,得到词和这些词的词性.将这些词组成基于词的词典用以提取词特征,其过程与第 1.2 节中提到的提取字特征的过程相似;
- 然后,我们将词特征、词性特征、实体指称特征、准确的实体类型特征和实体相对位置特征作为关系实例的组合特征,与任务 1 中相同的过程求得 DBN^i,SVM 和 NN 这 3 种机器学习模型进行关系抽取任务的结果.

我们将基于字特征和基于词特征的关系抽取结果进行比较,证明字特征更适用于关系抽取.

对于任务 3,我们设计了 3 组不同的实验,分别利用正确的实体类型特征、带噪音的实体类型特征和不利用实体类型特征,将它们分别组合实体指称特征、准确的实体类型特征和实体相对位置特征.利用 DBN^i,SVM 和 NN 这 3 种机器学习模型基于这 3 组特征进行关系抽取,验证实体类型对关系抽取的影响.

3.2 实验结果及分析

3.2.1 基于 DBN 模型的有效性验证

① 只识别关系实例

我们将包含语义关系的实例归为一类,相对地,将不包含语义关系的实例归为另一类.我们尝试了不同的 DBN 网络结构,各层隐含层的节点数自底向上依次为 2 700,2 100 和 1 800 是最佳的.神经网络节点数的增多,能够提高网络的逼近能力,但是会降低网络的泛化能力.所以节点数应逐层降维,提高网络泛化能力.降维幅度不宜过大,否则会丢失重要信息,并且会导致结果剧烈震荡[19].由于实例的组合特征是非线性的,实验证明,SVM 利用多项式核函数比线性核函数、RBF 核函数的效果都更好.我们尝试了不同的 SVM 参数,选取了其中最佳的参数(γ=3,$coef$=1.3).NN 的结构与 DBN 的结构一致,有利于比较.结果见表 2.

Table 2 Performances for detection only

表 2 只识别关系实例的结果

模型	准确率(%)	召回率(%)	F 系数(%)
DBN^1	73.18	70.72	71.93
DBN^2	75.51	69.99	72.64
DBN^3	75.86	70.86	73.28
SVM	77.99	66.91	72.02
NN	74.59	67.94	71.11

从结果中可以看出:虽然 DBN 未能取得 SVM 的准确率,但是召回率高于 SVM;DBN^1 的 F 系数不如 SVM,但是增加 1 层 RBM 后的 DBN^2 效果已经超过 SVM;最后,DBN^3 的召回率比 SVM 提高了 4%,F 系数提高了 1.2%,其效果要优于 SVM.DBN 与 NN 相比,3 项指标均有所提高,证明 RBM 是非常有效的自训练过程.表 3 是测试语料预定义的 5 类关系实例经过 DBN 模型得到的分类结果(给出准确率,便于对比),“+”代表实例被识别为包含语义关系,“–”表示实例被识别为不包含语义关系.

Table 3 Distribution of detection on each relation type by DBN

表 3 各类实例采用 DBN 模型识别的结果分布

类别	+	–	准确率(%)
Role	221	72	75.43
Part	112	28	80
At	139	78	64.06
Social	9	17	34.62
Near	3	4	42.86

从表 3 的结果可以看出:总的来说,数据的不平衡性导致占比例较大的关系类型的结果好于占比例较小的

关系类型;但是 At 类型的比例大于 Part 类型,结果却要更差.这是因为 At 类型实例内的两个实体,它们的类型组合比 Part 类型实体的类型组合更加复杂.同时,At 类型的大多数实例的实体相对位置是“分离型”的,与绝大多数不含语义关系的实例相似,使得 DBN 分类器容易将 At 类型分类错误.

② 先识别关系实例再识别实体类型

我们对上一组实验中被分类器识别为含有语义关系的实例进行细分,识别它们的关系类型.DBN 和 NN 的网络结构与第 1 组实验相同,SVM 依然采用多项式核函数,此时,γ=1 和 *coef*=1.3.结果见表 4.

Table 4 Performances for detection and classification in sequence

表 4 先识别再分类的结果

模型	准确率(%)	召回率(%)	*F* 系数(%)
DBN^1	68.17	63.98	66.01
DBN^2	69.36	64.16	66.69
DBN^3	69.79	64.28	66.92
SVM	72.09	61.64	66.46
NN	68.48	60.46	64.22

结果显示:DBN 依然在 3 项指标上超越 NN;DBN 在准确率上略微占优,在召回率上的改进比较明显.与 SVM 相比,DBN 未能取得相当的准确率,但是召回率依然有明显提高.其中,DBN^1 的 *F* 系数略差于 SVM,但是从 DBN^2 已经开始超越 SVM,DBN^3 的 *F* 系数较 SVM 略微提高 0.5%,效果提高不明显.这种先检测关系实力后识别关系类型的策略,3 种机器学习方法的效果差距不大,原因是,针对 ACE04 的语料而言,在关系类型识别阶段,训练语料的规模较小,只有 3 000 多个.

每一种类型的关系实例通过 DBN 分类器的分类结果见表 5.表 5 的结果与表 3 类似,数据的不平衡性使得每一类关系的分类结果存在差异,大致上是数据比重越大,效果越好.At 类型的效果不如 Part 类型,是因为 At 类型的特征组合的复杂度更高,导致更难分类.

Table 5 Performances of DBN for detection and classification in sequence on each relation type

表 5 各个类别关系实例采用 DBN 分类器的分类结果

类型	准确率(%)	召回率(%)	*F* 系数(%)
Role	73.96	72.7	73.32
Part	71.01	70	70.5
At	62.78	52.07	56.93
Social	72.73	30.77	43.24
Near	40	28.57	33.33

③ 直接识别关系的类型

在这一组实验中,DBN 和 NN 的结构依然不变,SVM 依然采用多项式核函数,其中,γ=2 和 *coef*=1.3.结果见表 6.实验证明,随着 DBN 层数的增加,分类效果越来越好.文献[13]已证明,3 层 RBM 已经足够取得良好的效果,所以本文实验中的 DBN 最多采用 3 层 RBM.采用 1 层 RBM 的 DBN 效果已经明显优于 NN,但比 SVM 略差;采用 2 层 RBM 的 DBN 与 SVM 的效果相当;采用 3 层 RBM 的 DBN 效果好于 SVM.以上 3 组实验均证明,DBN 是一种适合于关系抽取的机器学习模型,对稀疏特征向量有很好的分类能力,其效果优于传统的 SVM 和 NN.

Table 6 Performances for detection and classification in combination

表 6 直接识别关系类型的分类结果

模型	准确率(%)	召回率(%)	*F* 系数(%)
DBN^1	69.77	62.52	65.95
DBN^2	71.66	64.42	67.85
DBN^3	73.22	64.86	68.79
SVM	72.91	62.22	67.57
NN	69.13	56.07	61.93

表 7 给出了每一种类型的关系实例通过 DBN^3 分类器直接分类的结果,与表 5 的对比结果表明,直接分类的

效果比先识别再分类的效果要好.因为被识别为包含语义关系的实例的数据集合较小,使得在分类阶段分类器没有足够多的数据进行训练,导致结果较差.

从各组实验结果可以看出,相对于 NN,DBN 在准确率上略有提高,在召回率上的改进比较明显.相对于 SVM,在前两组实验中,DBN 的准确率不如 SVM,但是召回率的提高幅度比较大.在第 3 组实验中,深层的 DBN 在准确率和召回率两项指标上都超过了 SVM.实验证明,任何一组实验中的深层 DBN 的效果均好于 SVM 和 NN,DBN 非常适用于关系抽取任务.

Table 7 Performances of DBN for detection and classification in combination on each relation type

表 7 各个关系类型通过 DBN 分类器的直接分类结果

类型	准确率(%)	召回率(%)	*F* 系数(%)
Role	78.81	72.35	75.44
Part	72.97	77.14	75
At	65.7	52.07	58.1
Social	88.89	30.77	45.71
Near	28.57	28.57	28.57

3.2.2 基于字特征的有效性验证

在传统的中文信息抽取任务中,由于中文词法、句法和语法的特殊性,通常先进行中文分词,然后采用词和词性作为组合特征之一.但是中文分词系统依然存在误差,使用词特征会将噪音引入关系抽取任务中.为此,本文采用了基于字特征的方法,哪些字将组成一个中文词、哪些字不会同时出现在同一个中文词里,机器学习模型通过对训练语料的学习获得知识.

为了验证基于字特征的方法比基于词特征的方法更有效、更具有可分类性,我们将任务(1)中的字特征替换为词特征和词性特征的组合,比较基于字特征和基于词特征的优劣.基于字特征的字典的维数为 1 498,基于词特征的字典的维数是 3 754,词性特征的维数是 74.基于词特征的特征组合将极大地增加特征向量的复杂度.由于任务(1)的结果,直接识别关系实例类型的方式比先识别关系实例再识别关系类型的方式效果更好,所以任务(2)采用直接识别关系实例的方式比较字特征和词特征的优劣,实验结果见表 8.

Table 8 Performances for detection and classification in combination with word-based feature

表 8 基于词特征的关系实例直接分类结果

模型	特征	准确率(%)	召回率(%)	*F* 系数(%)
DBN^3	字	73.22	64.86	68.79
DBN^3	词	72.01	54.61	62.11
DBN^3	词+词性	66.02	59.24	62.45
DBN^2	字	69.36	64.16	66.69
DBN^2	词	71.97	54.56	62.07
DBN^2	词+词性	65.91	59.15	62.35
DBN^1	字	69.77	62.52	65.95
DBN^1	词	70.15	54.03	61.04
DBN^1	词+词性	66.26	56.08	60.75
SVM	字	72.91	62.22	67.57
SVM	词	68.73	56.6	62.07
SVM	词+词性	68.67	56.80	62.2
NN	字	69.13	56.07	61.93
NN	词	66.53	48.9	56.37
NN	词+词性	65.7	49.63	56.55

我们将字特征替换为两组特征,第 1 组是词特征,第 2 组是词特征加词性特征.其实验结果显示,用这两组特征替换字特征都达不到字特征的效果.基于字特征与只用词特征的结果相比,两者的准确率相差不大,而字特征较大地提高了召回率.证明基于字的特征比基于词的特征更适合于关系抽取任务,说明基于词的特征的复杂性反而降低了组合特征的可分类性;并且,中文分词系统和中文词性标注系统的结果都有一定的误差,会将噪音带入组合特征之中,从而影响了关系抽取的效果.我们再比较基于词特征的两组组合特征,加入了词性特征并没有

为关系抽取结果带来明显的 F 系数的提高.同时,增加了词性特征之后提高了关系抽取的召回率,但是降低了准确率.随着组合特征的增多,关系抽取的准确率会下降,召回率会升高[20].

3.2.3 实体类型信息对关系抽取的影响

我们设计了 3 组不同的实验,分别利用正确的实体类型特征、带噪音的实体类型特征和不利用实体类型特征验证实体类型对关系抽取的影响.

为了得到带噪音的实体类型特征,我们首先利用 DBN,SVM 和 NN 分别建立实体类型分类器.我们利用 4 折交叉验证,假设所有实体的边界信息已经确定,提取了语料中 7 746 个实体作为训练语料,2 482 个实体作为测试语料.组合特征为字特征和指称特征.DBN 的隐含层的节点数自底向上依次为 900,600,300;SVM 利用了多项式核函数(γ=1.2,*coef*=1.1).由于实验只是对已识别边界的实体进行分类,其召回率等于准确率,结果见表 9.结果表明,DBN 的效果依然优于 SVM 和 NN,证明其同样适合于实体分类任务.

Table 9 Performances of entity classification

表 9 实体类别的分类结果

模型	准确率(%)
DBN^3	91.45
DBN^2	91.42
DBN^1	91.05
SVM	90.82
NN	87.23

表 10 给出了 3 组利用不同实体特征的关系抽取结果,模型名称的下标代表所使用的实体类型特征,c 代表正确的实体类型特征,p 代表利用通过实体类型分类器得到的有噪音的实体类型特征,n 代表不使用实体类型特征.所有模型的参数与表 6 中实验的参数相同,并采用对关系实例直接分类方式进行关系抽取.

Table 10 Performances of combining different entity type feature

表 10 利用不同实体类型特征的关系抽取结果

模型	准确率(%)	召回率(%)	F 系数(%)
DBN_c	73.22	64.86	68.79
DBN_p	66.94	59.88	63.21
DBN_n	64.7	54.47	59.14
SVM_c	72.91	62.22	67.57
SVM_p	71.22	57.24	63.47
SVM_n	68.21	53.73	60.11
NN_c	69.13	56.07	61.93
NN_p	64.49	52.12	57.65
NN_n	61.5	45.39	52.23

实验结果表明,使用正确的实体类型特征信息比不使用实体类型的特征信息效果有明显的提高,证明实体类型是关系抽取的一个重要组合特征.使用有噪音的实体类型特征的结果,大致介于正确的和不使用实体类型特征结果的中间状态,即使实体类型分类器已经达到超过 90%准确率的效果,但是实体类型的噪音对关系抽取的影响依然较大.我们再进一步分析,利用有噪音的实体类型特征和不利用实体类型特征,SVM 的效果要略优于 DBN,说明正确的实体类型特征对 DBN 模型的帮助更大.

4 结论及展望

DBN 对于关系抽取是一种全新的机器学习算法,它对高维特征向量具有很强的提取特征和分类特征能力,其深层结构更能帮助它提取出更抽象、更具可分类性的特征.实验结果表明,DBN 在关系抽取任务中的效果好于 SVM 和 NN,是一种非常适用于信息抽取领域的算法.我们对比了基于字特征和基于词特征的优劣,证明,在关系抽取任务中,字特征是更加适合表示关系实例的特征.最后,我们还对比了正确的实体类型特征、带噪音的实体类型特征和不使用实体类型特征对关系抽取效果的影响.将来的工作拟在以下几个方面展开:

(1) 将本文提出的方法在更大规模的中文数据集 ACED05 以及英语数据集上测试,以进一步验证方法的

有效性,进一步的测试需要更大规模的实验平台;

(2) 利用该方法多任务地进行实体抽取与关系抽取.

References:

[1] He TT, Xu C, Li J, Zhao JZ. Named entity relation extraction method dased on seed self-expansion. Compter Engineering, 2006,32(21):183−184, 193 (in Chinese with English abstract).

[2] Fan N, Cai WD, Zhao Y. Extraction of subjective relation in opinion sentences based on maximum entropy model. Compter Engineering, 2010,36(2):4−6 (in Chinese with English abstract).

[3] Che WX, Liu T, Li S. Automatic entity relation extraction. Journal of Chinese Information Processing, 2005,19(2):1−6 (in Chinese with English abstract).

[4] Huang X, Zhu QM, Qian LH, Liu MM. Chinese entity relation extraction based on features combination. Microelectronics & Computer, 2010,27(4):198−200, 204 (in Chinese with English abstract).

[5] Liu KB, Li F, Liu L, Han Y. Implementation of a kernel-based Chinese relation extraction system. Journal of Computer Research and Development, 2007,44(8):1406−1411 (in Chinese with English abstract).

[6] Culotta A, Sorensen J. Dependency tree kernel for relation extraction. In: Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004). Barcelona, 2004. 423−429. [doi: 10.3115/1218955.1219009]

[7] Bunescu RC, Mooney RJ. A shortest path dependency kernel for relation extraction. In: Proc. of the Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing. Vancouver, 2005. 724−731. [doi: 10.3115/1220575.1220666]

[8] Zhang M, Zhang J, Su J. Exploring syntactic features for relation extraction using a convolution tree kernel. In: Proc. of the Human Language Technology Conf. of the North American Chapter of the Association for Computational Linguistics. New York: Springer-Verlag, 2006. 288−295. [doi: 10.3115/1220835.1220872]

[9] Zhao J, Wang XL, Guan Y. Comparing feature combination with features fusion in Chinese named entity recognition. Journal of Computer Applications, 2005,25(11):2647−2649 (in Chinese with English abstract).

[10] Jing HY, Florian R, Luo XQ, Zhang T, Ittycheriah A. How to get a Chinese name (entity): Segmentation and combination issues. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Sapporo, 2003. 200−207.

[11] Giuliano C, Lavelli A, Romano L. Relation extraction and the influence of automatic named-entity recognition. ACM Trans. on Speech and Language Processing (TSLP), 2007,5(1):1−26. [doi: 10.1145/1322391.1322393]

[12] Zhang P, Li WJ, Wei FR, Lu Q, Hou YX. Exploiting the role of position feature in Chinese relation extraction. In: Proc. of the 6th Int'l Conf. on Language Resources and Evaluation (LREC). Marrakech, 2008. 28−30.

[13] Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18:1527−1554. [doi: 10.1162/neco.2006.18.7.1527]

[14] Bengio Y, LeCun Y. Scaling learning algorithms towards AI. In: Proc. of the Large-Scale Kernel Machines. MIT Press, 2007. http://www.iro.umontreal.ca/~lisa/bib/pub_subject/language/pointeurs/bengio+lecun-chapter2007.pdf

[15] Hinton GE. Products of experts. In: Proc. of the 9th Int'l Conf. on Artificial Neural Networks (ICANN), Vol.1. Edinburgh, 1999. 1−6.

[16] Neal RM. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report, CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993. http://www.cs.toronto.edu/~radford/review.abstract.html

[17] Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Computation, 2002,14(8):1771−1800. [doi: 10.1162/089976602760128018]

[18] Carreira-Perpinan MA, Hinton GE. On contrastive divergence learning. In: Proc. of the Artificial Intelligence and Statistics (AISTATS 2005). Barbados, 2005. http://learning.cs.toronto.edu/~hinton/absps/cdmiguel.pdf

[19] Xia KW, Li CB, Shen JY. An optimization algorithm on the number of hidden layer nodes in feed-forward neural network. Computer Science, 2005,32(10):143−145 (in Chinese with English abstract).

[20] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proc. of the Association for Computational Linguistics 2004 on Interactive Poster and Demonstration Sessions. Barcelona, 2004. [doi: 10.3115/1219044.1219066]

附中文参考文献:

[1] 何婷婷,徐超,李晶,赵君喆.基于种子自扩展的命名实体关系抽取方法.计算机工程,2006,32(21):183−184,193.

[2] 樊娜,蔡皖东,赵煜.基于最大熵模型的观点句主观关系提取.计算机工程,2010,36(2):4−6.

[3] 车万翔,刘挺,李生.实体关系自动抽取.中文信息学报,2005,19(2):1−6.

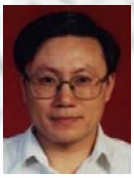
[4] 黄鑫,朱巧明,钱龙华,刘梅梅.基于特征组合的中文实体关系抽取.微电子学与计算机,2010,27(4):198−200,204.

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

[9] 赵健,王晓龙,关毅.中文名实体识别中的特征组合与特征融合的比较.计算机应用,2005,25(11):2647−2649.

[19] 夏克文,李昌彪,沈钧毅.前向神经网络隐含层节点数的一种优化算法.计算机科学,2005,32(10):143−145.

陈宇(1983－),男,广东五华人,博士生,主要研究领域为信息抽取,自然语言处理.



赵铁军(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算语言学,人工智能,机器翻译.

郑德权(1968－),男,博士,副教授,CCF 会员,主要研究领域为自然语言处理,信息抽取,机器学习.