基于深度信念网络的命名实体识别

冯蕴天 张宏军 郝文宁 陈 刚

(解放军理工大学指挥信息系统学院 南京 210007)

摘 要 传统的命名实体识别方法是将大量手工制定的特征输入到统计学习模型中以实现对词语的标记,能够取得较好的效果,但其手工特征制定的方式增加了模型建立的难度。为了减轻传统方法中手工特征制定的工作量,首先对神经网络语言模型进行无监督训练以得到词语特征的分布式表示,然后将分布式的特征输入到深度信念网络中以发现词语的深层特征,最后进行命名实体识别。该方法在前人研究的基础上利用深度信念网络对神经网络语言模型进行了扩展,提出了一种可用于命名实体识别的深层架构。实验表明,在仅使用词特征和词性特征的条件下,该方法用于命名实体识别的性能略优于基于条件随机场模型的方法,具有一定的使用价值。

关键词 深度信念网络,命名实体识别,神经网络语言模型

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.046

Named Entity Recognition Based on Deep Belief Net

FENG Yun-tian ZHANG Hong-jun HAO Wen-ning CHEN Gang

(Institute of Command Information System, PLA University of Science and Technology, Nanjing 210007, China)

Abstract Traditional named entity recognition methods, which tag words by inputting a good deal of handmade features into statistics learning models, have achieved good results, but the manual mode of defining features makes it more difficult to build the model. To decrease the workload of the manual mode, this paper firstly got the distributed representation of word features by training the neural network language model without supervision, then discovered the deep features of words by inputting the distributed features into the deep belief net, finally conducted named entity recognition. The method uses the deep belief net to extend the neural network language model on the basis of research of predecessors, and presents a deep architecture which is available for named entity recognition. Experiments show that the method applied to named entity recognition can perform better than traditional conditional random field model if both only using term feature and POS feature, and has a certain use value.

Keywords Deep belief net, Named entity recognition, Neural network language model

1 引言

命名实体(Named Entity, NE)[1]是文本中基本的信息单位,主要包括固有名称、缩写和其他唯一标识等,是正确理解文本的基础。命名实体识别(Named Entity Recognition, NER)[2]是信息提取、问答系统、句法分析、机器翻译和面向Semantic Web的元数据标注等应用领域的重要基础性工作,在自然语言处理技术走向实用化的过程中占有重要地位。

目前,最常用的命名实体识别方法为基于条件随机场模型的方法。条件随机场模型首先被应用于解决英文命名实体识别问题[2],取得了良好的效果。2006年王志强等[3]将条件随机场模型应用于中文命名实体识别任务中,进一步提高了中文命名实体识别系统的性能。基于条件随机场模型的方法主要是依据逻辑直觉和训练语料中的统计信息手工制定出大量的特征并输入模型,其识别性能很大程度上依赖于特征的质量,因此会产生以下两点不足:手工制定的特征数量通常较大,导致产生的结果模型过大,会增加应用过程中计算和存储的消耗;手工特征制定的方式需要反复进行实验并修正,工作量较大,且要求研究者具备大量的语言学知识。

2006年,加拿大多伦多大学的Hinton教授[4]提出了深度学习的概念,在机器学习领域掀起了一次热潮。深度学习是通过模仿人脑的多层抽象机制来实现对数据(图像、语音和文本等)的抽象表达,将特征学习和分类整合到一个统一的学习框架中,减少了手工特征制定的工作量。近年来,深度学习在图像识别和语音识别等领域已经取得了很大的成功,深度学习框架在自然语言处理领域同样也受到了很多关注。由于深度学习技术可以在原始字符集上提取出高级特征,因此本文使用神经网络语言模型在大量未标记语料上无监督地学习出词特征和词性特征的分布式表示,以取代传统模型中的特征表示方式,并将分布式的特征输入到深度信念网络中,其中

到稿日期:2015-02-01 返修日期:2015-05-14

冯蕴天(1990—),男,硕士生,主要研究方向为军用数据与知识工程,E-mail:fengyuntian2009@live.cn;**张宏军**(1963—),男,教授,博士生导师,主要研究方向为军事建模与仿真;**郝文宁**(1971—),男,副教授,主要研究方向为军用数据与知识工程;**陈 刚**(1974—),男,副教授,主要研究方向为数据工程。

的隐单元被训练来捕捉原始特征内部的相关性，可以最大程度地减轻手工特征制定的工作量，最终实现对词语的标记。

2 深度信念网络

深度学习通过学习一种深层非线性网络结构，实现复杂函数逼近，表征输入数据的分布式表示，学习出更高级的特征，从而最终提升分类或预测的准确性。其本质是构建出具有多个隐层的学习模型，常用的模型包括自动编码器、深度信念网络和卷积神经网络等。这些模型也是可自动学习出海量数据潜在分布的多层表达算法。其中最常用、最经典的模型是深度信念网络。

2006 年 Hinton 等[5]提出了一种新的贪婪逐层无监督算法来初始化基于受限玻尔兹曼机的深度信念网络。深度信念网络(Deep Belief Net，DBN)是一种无监督的深度学习结构，是由一系列受限玻尔兹曼机(RBM)单元组成。深度信念网络作为经典的深度学习方法，包含了较多隐藏层，可以更好地学习各种复杂数据的结构和分布。

2.1 受限玻尔兹曼机

假设有一个包含两层的无向图模型，每一层的节点之间没有连接。一层 v 是可见层，表示数据的输入；一层 h 是隐藏层，表示特征的提取。可见层 v 包含 m 个可见单元，隐藏层 h 包含 n 个隐藏单元，且所有的可见单元和隐藏单元通常都是随机二值变量节点(只能取值 0 或者 1)，其分布满足伯努利分布，则称这个模型是受限玻尔兹曼机(Restricted Boltzmann Machine，RBM)。RBM 的无向图模型如图 1 所示。

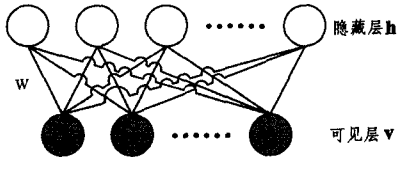

图 1 RBM 的无向图模型

假设可见单元的状态用向量 $\boldsymbol{v}$ 表示，隐藏单元的状态用向量 $\boldsymbol{h}$ 表示，且 $\boldsymbol{v}$ 和 $\boldsymbol{h}$ 同时满足伯努利分布，则可见层节点和隐藏层节点$(\boldsymbol{v},\boldsymbol{h})$之间的能量函数为

$$E(\boldsymbol{v},\boldsymbol{h}|\theta)=-\sum_{i=1}^{m}\sum_{j=1}^{n}w_{ij}v_ih_j-\sum_{i=1}^{m}a_iv_i-\sum_{j=1}^{n}b_jh_j$$

其中，$\theta=\{w_{ij},a_i,b_j\}$ 是 RBM 的参数，w_{ij} 表示可见单元与隐藏单元之间的连接权重，a_i 表示可见单元 i 的偏置，b_j 表示隐藏单元 j 的偏置，v_i 和 h_j 表示可见单元 i 和隐藏单元 j 的二值状态。

RBM 学习的目的是得到参数 θ，找到系统的最小全局能量结构，将能量函数指数化并且正则化，可以得到可见层节点和隐藏层节点$(\boldsymbol{v},\boldsymbol{h})$的联合概率分布

$$P(\boldsymbol{v},\boldsymbol{h}|\theta)=\frac{e^{-E(\boldsymbol{v},\boldsymbol{h}|\theta)}}{Z(\theta)}$$

其中，

$$Z(\theta)=\sum_{\boldsymbol{v},\boldsymbol{h}}e^{-E(\boldsymbol{v},\boldsymbol{h}|\theta)}$$

式中，$Z(\theta)$为归一化因子。

2.2 深度信念网络结构

深度信念网络是一种包含多个受限玻尔兹曼机的深层结构，作为一种特征提取方法，可通过组合低层特征形成更加抽象的高层表示，其网络结构如图 2 所示。DBN 增加了中间 RBM 网络的层数，自底向上用每一层 RBM 对输入数据进行提取、抽象，尽可能地保留重要信息。RBM 有效的学习过程使它适于作为 DBN 的构成模块，将最后一层 RBM 的输出信息作为 BP 神经网络(最终分类器)的输入数据。因此，DBN 也可以认为是带有训练初始权值的神经网络。

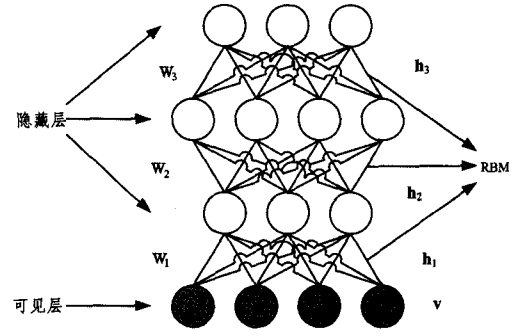

图 2 深度信念网络结构

3 构建分布式特征

3.1 分布式表示

执行命名实体识别任务一般可分为文本预处理、特征的提取与表示、模型的训练、命名实体识别的实施等步骤。在确定使用的统计学习模型后，所构建特征的质量好坏将直接决定着命名实体识别系统的性能。目前，命名实体识别系统中最常用的词语特征表示方法是把每个词语特征表示为一个很长的向量，其维度为该种词语特征的总数量，若采用稀疏的方式进行存储，则可以给每个词语特征分配一个数字 ID，这种简洁的表示方法与支持向量机、最大熵和条件随机场等模型相结合已经被广泛地用于各种自然语言处理任务中。

在命名实体识别任务中，通常将经过文本预处理后的每个词语本身作为一种特征，称为词特征，词特征能够较完整地反映原始字符集的基本信息。若训练语料规模足够大，仅使用词特征就能较好地完成命名实体识别任务。除此之外，大量研究表明[6]把词性作为一种特征输入模型可显著地提高命名实体识别系统的性能，称为词性特征。本文使用的标准汉语词性标记集中共有 99 个标记，其中包括 22 个一类标记、66 个二类标记和 11 个三类标记。因此，为了最大程度地减少手工特征制定的工作量，本文仅提取并使用最原始的词特征和词性特征用于命名实体识别任务。

1986 年 Hinton 等[7]将分布式表示的思想引入到符号数据中，对符号数据进行分布式表示是神经网络模型最核心的思想之一。2013 年 Wang 等[8]对多种非线性模型进行了经验性的研究，结果表明非线性模型通常是高效的，主要是由于采用了分布式表示。若对词特征和词性特征进行传统的特征表示，那么任意两个词语之间或者任意两个词性标记之间都是孤立的、没有联系的。大量研究表明[9]，对词语的特征进行分布式表示可以有效改进自然语言处理系统的性能。对词特征和词性特征进行分布式表示，即把每个词语或者每个词性标记都表示为一个低维实数向量，那么任意两个词语之间或者任意两个词性标记之间的欧氏距离将更近。词语特征的分布式表示可解决机器学习中的维数灾难和局部泛化限制等问题，相比于传统的特征表示方式可以更深入地探索输入数据之间的固有联系，捕获其内部的语法、语义相似性。当遇到训练语料中未出现的词语或词性标记时，采用词语特征的分布式表示训练出的模型仍然能够有较好的表现。因此，本文将

在大量未标记语料上无监督地学习出词特征和词性特征的分布式表示，并输入到深度信念网络中。

3.2 神经网络语言模型

神经网络语言模型(NNLM)是 Bengio 等[10]于 2003 年提出的，与传统的 n-gram 语言模型不同的是，神经网络语言模型可从大量未标记数据中学习出词语的分布式表示，并且可以对词语之间的关系以及相似度进行建模。近年来，神经网络语言模型开始逐渐被应用于多种自然语言处理任务中，并取得了良好的效果。2011 年 Mikolov 等[11]使用循环神经网络改进了 Bengio 的神经网络语言模型，其应用于语音识别上的性能要优于传统的 n-gram 语言模型。2011 年 Collobert 等[12]提出了一个统一的神经网络架构及其学习算法，并设计了 SENNA 系统可用于解决语言建模、词性标记、组块分析、命名实体识别、语义角色标记和句法分析等问题。2013 年 Zheng 等[13]在大规模未标记数据集上改进了中文词语的内在表示形式，并使用深度学习模型发现词语的深层特征以解决中文分词和词性标记问题。2013 年 Mansur 等[9]采用了一种基于特征的神经网络语言模型，其可结合上下文特征估计出词语出现的可能性以解决中文分词问题。2014 年 Pei 等[14]提出一种新奇的神经网络语言模型并应用于中文分词任务，使用标记嵌入和基于张量的变换对词语特征之间的关系进行建模。

神经网络语言模型指利用神经网络对各种语言单位(字、词、句子和整篇文章等)进行概率分布的估计，该概率分布是一种语言的生成模型，可以捕获自然语言的统计规律以改善各种自然语言处理系统的性能。因此，本文将在前人研究的基础上，利用神经网络语言模型寻找词语特征在连续向量空间中的表示，构建出分布式特征。

3.3 无监督训练

本文构建的神经网络语言模型如图 3 所示。模型的求解目标为 $P(F_t|F_c)$，该条件概率表示考虑当前词语的上下文语境 F_c 时，当前词语的特征为 F_t 的概率。词语的分布式特征记为 $F_1,F_2,\cdots,F_V$，该种分布式特征的总数量为 V，其中的每个特征都用一个 m 维的实值列向量表示。

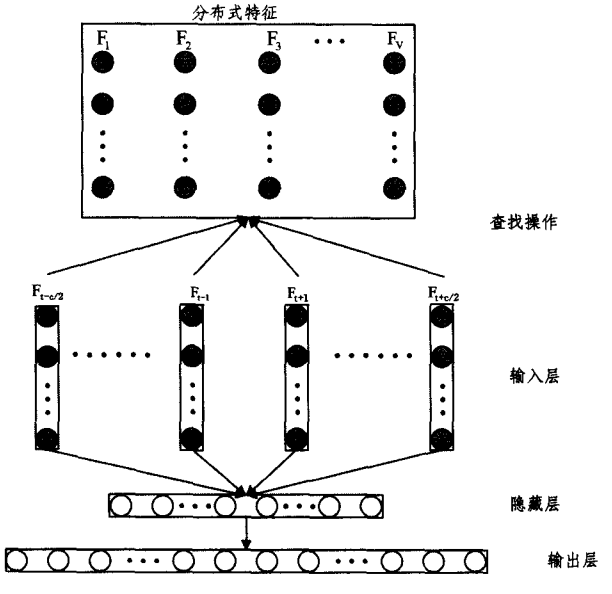

图 3 神经网络语言模型

第 1 层(查找操作)：列向量 $F_1,F_2,\cdots,F_V$ 内的元素为该模型中待训练的参数，可在这些列向量中查找出当前词语上下文的分布式特征，此时每个特征都是一个 m 维的实值列向量。

第 2 层(输入层)：将当前词语 t 上下文语境中的 c 个实值列向量进行首尾拼接组成一个 $c\times m$ 维的特征向量 x，并把该特征向量 x 作为神经网络的输入。

第 3 层(隐藏层)：进行非线性变换 $Hx+d$，并用 $tanh$ 激活函数使 $z=tanh(Hx+d)$，输入为特征向量 x，$H\in\mathbb{R}^{(c\times m)\times h}$ 为变换矩阵，$d\in\mathbb{R}^h$ 为偏置矩阵，h 为隐藏层内节点的个数，经过隐藏层后，$c\times m$ 维的特征向量 x 变成为了 h 维的向量 z。

第 4 层(输出层)：进行非线性变换 $Wz+b$，并用 $softmax$ 函数做概率归一化，输入为 h 维的向量 z，$W\in\mathbb{R}^{h\times V}$ 为变换矩阵，$b\in\mathbb{R}^V$ 为偏置矩阵，经过 $Wz+b$ 后产生了 V 个没有归一化的值，对其做 $softmax$ 可得到 $P(F_t|F_c)$，其中第 i 个输出值为 $P(F_i|F_c)$，表示当前词语的特征为 F_i 的概率。

该模型中的未知参数 θ 包括变换矩阵 H 和 W、偏置矩阵 d 和 b，以及词语的分布式特征 $F_1,F_2,\cdots,F_V$，本文使用后向传播算法无监督地训练该神经网络语言模型。$F_1,F_2,\cdots,F_V$ 从 V 组随机的初始值开始，对于每一个训练样本 (F_c,F_t)，模型的目标是最大化 $P(F_t|F_c)$。对于所有训练样本，训练的准则为使最大似然函数 $-\sum_{t=1}^{T}\log P(F_t|F_{t-\frac{c}{2}},\cdots,F_{t-1},F_{t+1},\cdots,F_{t+\frac{c}{2}})+\lambda$ 取最小，其中 T 为训练样本集，λ 为正则项。具体求解过程中采用了随机梯度下降算法，梯度更新的准则为 $\theta\leftarrow\theta+\varepsilon\frac{\partial\log p(F_t|F_{t-\frac{c}{2}},\cdots,F_{t-1},F_{t+1},\cdots,F_{t+\frac{c}{2}})}{\partial\theta}$，其中 ε 为学习率，通常取值为 0.02。

本文首先从训练语料中提取出词语序列样本集和词性序列样本集，然后分别在两种样本集上无监督地训练上述神经网络语言模型，即可获得词特征和词性特征的分布式表示。词特征的分布式表示为 $Fw_1,Fw_2,\cdots,Fw_{V_1}$ 等 V_1 个列向量，V_1 等于词语序列样本集中不同词语的总数量；词性特征的分布式表示为 Fp_1、Fp_2、$\cdots$、Fp_{V_2} 等 V_2 个列向量，V_2 等于词性序列样本集中不同词性的总数量。本文把词特征的分布式表示的维数设定为 50，把词性特征的分布式表示的维数设定为 25。

4 一种可用于命名实体识别的深层架构

命名实体识别本质上可视作序列数据的标记问题，即为每一个词语分配一个标记以表示其所属的类别。在进行命名实体识别时，研究者的主要精力通常集中在特征的制定上，特征的制定需要耗费大量的人力、物力，并且需要依靠逻辑直觉和大量的语言学知识。为了减少对手工特征制定的依赖，本文建立了一种可用于命名实体识别的深层架构，该架构的本质是构建具有很多隐层的深度信念网络，以学习出更有用的特征，从而提升识别的性能。相比于自然语言处理任务中常用的条件随机场模型，该架构具有两大优势：1)传统的稀疏特征被稠密的分布式特征所取代；2)深度学习结构被使用以发现更高级的特征。

4.1 深层架构

2011 年 Collobert 等[12]重构了神经网络模型并应用于多种自然语言处理任务中。本文在 Collobert 工作的基础上，使

用深度信念网络建立了一种深层架构，利用多个隐层从输入的原始分布式特征中发现更多的深层特征，该深层架构如图4所示。

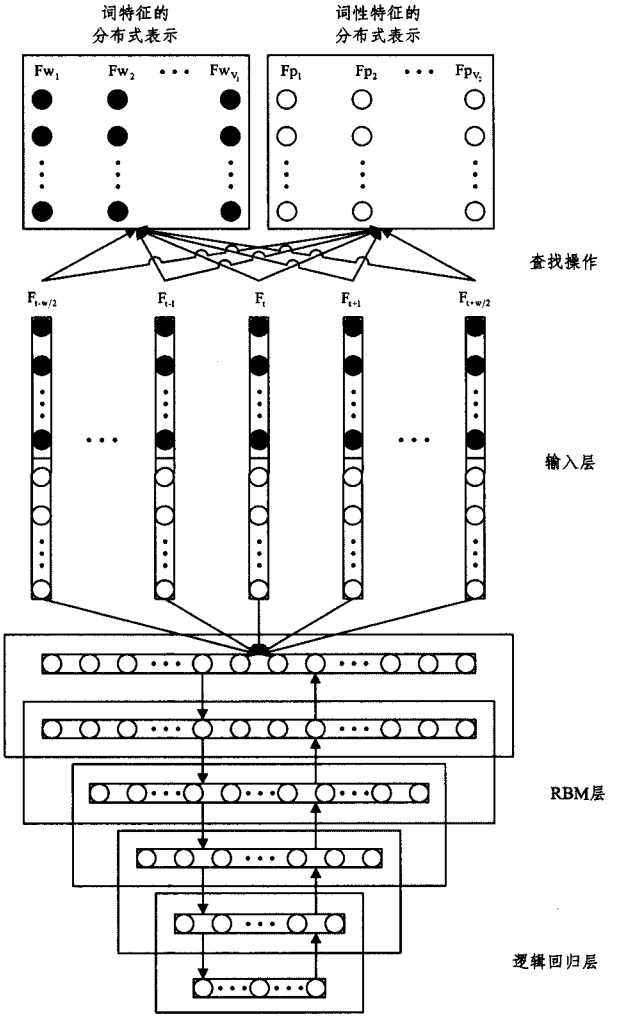

图4 可用于命名实体识别的深层架构

给定一个中文句子 $c_{[1:n]}$，表示一个包含 n 个词语的序列，本文使用一个大小为 w 的活动窗口划过整个句子，从词语 c_1 到 c_n。窗口模式是自然语言处理任务中最常用的一种模式，假设当前词语 c_t 的标记主要取决于其附近词语，被用于控制上下文的长度。若活动窗口过大，则会向模型中引入过多噪声，故本文将活动窗口的大小 w 设置为5。

第1层（查找操作）：已知词特征的分布式表示 Fw_1，Fw_2，…，Fw_{V_1} 和词性特征的分布式表示 Fp_1，Fp_2，…，Fp_{V_2}，在大小为 w 的活动窗口内收集当前词语 c_t 上下文的特征，并在上述列向量中查找出这些特征的分布式表示，将活动窗口中每个词语对应的词特征的分布式表示和词性特征的分布式表示首尾相连，形成了特征向量 $F_{t-\frac{w}{2}}$，…，F_{t-1}，F_t，F_{t+1}，…，$F_{t+\frac{w}{2}}$，实现了将活动窗口内每一个词语的词特征和词性特征投射为一个特征向量，每一个特征向量的维数为75。

第2层（输入层）：将当前词语 c_t 的活动窗口内的特征向量 $F_{t-\frac{w}{2}}$，…，F_{t-1}，F_t，F_{t+1}，…，$F_{t+\frac{w}{2}}$ 进行首尾拼接组成一个5×75维的向量 x，并把该向量 x 作为深度信念网络的输入。

第3层（RBM层）：该层由若干个RBM串联而成。第一个RBM将输入的向量 x 视为可见层，并使用高斯分布对输入的向量 x 进行建模，假设其分布满足高斯分布，其节点个数同样为5×75，隐藏层包含了若干个随机二值变量节点，其分布满足伯努利分布。第二个RBM的可见层即为第一个RBM的隐藏层，其可见层和隐藏层的节点都满足伯努利分布。将此种建模和传递方法延续到最后一个RBM。每个RBM的可见层和隐藏层内节点个数的设定必须兼顾计算速度和模型的效果，其节点个数通常是逐层减少的，可理解为下一层能够保留上一层中的主要信息。本文将若干个RBM串联作为一种特征表示方法，可自动地发现有用的、更高级的特征，从而减少手工特征制定的工作量。

第4层（逻辑回归层）：该层使用了BP神经网络计算每一种标记的得分。进行非线性变换 $Wz+b$，并用 $softmax$ 函数做概率归一化，输入向量 $z\in\mathbb{R}^y$ 为最后一个RBM的隐藏层，$W\in\mathbb{R}^{y\times n}$ 为变换矩阵，$b\in\mathbb{R}^n$ 为偏置矩阵，n 为输出层的节点个数，等于所定义的标记集的大小，经过 $Wz+b$ 后产生了 n 个没有归一化的值，对其做 $softmax$ 可得到每一种标记的网络得分，其中第 i 个输出值为当前词语在第 i 个标记上的网络得分。

4.2 参数训练

对该深层架构的训练本质是在训练语料上计算模型中的未知参数，未知参数主要包括RBM层中若干个RBM的参数 $\theta=\{w_{ij},a_i,b_j\}$、逻辑回归层中的变换矩阵 $W\in\mathbb{R}^{y\times n}$ 和偏置矩阵 $b\in\mathbb{R}^n$。其训练过程主要分为以下两个步骤：

第1步：使用RBM的自训练算法对RBM层内的若干个RBM进行逐个训练，初始化若干个参数 $\theta=\{w_{ij},a_i,b_j\}$ 以实现对该深层架构的预训练。

通常将对RBM的参数 $\theta=\{w_{ij},a_i,b_j\}$ 的学习转化为对极大对数似然 $\log P(\boldsymbol{v}|\theta)$ 的估计，即

$$\theta=\arg_\theta\max\log P(\boldsymbol{v}|\theta)$$

其中，$\log P(\boldsymbol{v})=\log\sum_{\boldsymbol{h}}\exp[-E(\boldsymbol{v},\boldsymbol{h}|\theta)]-\log\sum_{\boldsymbol{v},\boldsymbol{h}}\exp[-E(\boldsymbol{v},\boldsymbol{h}|\theta)]$。

采用梯度下降法对对数似然求偏导，得到参数 θ 的学习法则：

$$\frac{\partial\log P(\boldsymbol{v}|\theta)}{\partial\theta}$$

$$=\frac{\sum_{\boldsymbol{h}}[\exp(-E(\boldsymbol{v},\boldsymbol{h}|\theta))(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta})]}{\sum_{\boldsymbol{h}}\exp(-E(\boldsymbol{v},\boldsymbol{h}|\theta))}-$$

$$\frac{\sum_{\boldsymbol{v},\boldsymbol{h}}[\exp(-E(\boldsymbol{v},\boldsymbol{h}|\theta))(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta})]}{\sum_{\boldsymbol{v},\boldsymbol{h}}\exp(-E(\boldsymbol{v},\boldsymbol{h}|\theta))}$$

$$=\sum_{\boldsymbol{h}}[P(\boldsymbol{h}|\boldsymbol{v})(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta})]-\sum_{\boldsymbol{v},\boldsymbol{h}}(P(\boldsymbol{v},\boldsymbol{h})(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta}))$$

$$=E_{P(\boldsymbol{h}|\boldsymbol{v})}(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta})-E_{P(\boldsymbol{v},\boldsymbol{h})}(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta})$$

其中，$E_{P(\boldsymbol{h}|\boldsymbol{v})}(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta})$ 表示偏导 $-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta}$ 在训练语料“先验分布”$P(\boldsymbol{h}|\boldsymbol{v})$ 下的期望，$E_{P(\boldsymbol{v},\boldsymbol{h})}(-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta})$ 表示 $-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta}$ 在联合分布 $P(\boldsymbol{v},\boldsymbol{h})$ 下的期望，$-\frac{\partial E(\boldsymbol{v},\boldsymbol{h}|\theta)}{\partial\theta}$ 根据RBM能量公式很容易计算。

上式中的第二项是联合概率分布的期望，由于归一化因子的存在，联合概率分布 $P(\boldsymbol{v},\boldsymbol{h})$ 很难通过直接计算获取，因

此无法直接、精确地计算上式的第二项。对此，较好的解决方法是利用 Gibbs 采样[15]来近似计算梯度。因此，可使用对比散度(Contrastive Divergence，CD)[16]算法获得 RBM 可见层节点和隐藏层节点之间的权值。CD 算法其实就是不断缩小真实数据分布与经 RBM 重构的数据分布之间差异的过程，并不断进行依赖逼近，将隐藏层激活单元(即可见层的重构单元)和可见层输入单元之间的相关性差别作为权值更新的主要依据，并运用 Gibbs 采样来进行迭代，直到其达到收敛，最终即可求得参数。

第 2 步：在标记语料上使用后向传播算法对整个深层架构性能进行优化调整，可获得逻辑回归层中的参数 $W\in\mathbb{R}^{y\times n}$ 和 $b\in\mathbb{R}^{n}$，并进一步更新上一步学习到的若干个 RBM 的参数 $\theta=\{w_{ij},a_i,b_j\}$，使整个深层架构达到最优。首先为前向传播，将最后一个 RBM 隐藏层的输出信息传递到逻辑回归层，把所提取的最高级特征映射到相应的标记信息上，利用数据的标记值对模型进行有监督训练，并不断调整连接权值，减小模型的目标预测标记与实际标记之间的误差；然后为反向传播，计算前向传播过程中目标预测标记与实际标记之间的误差，并将该误差从逻辑回归层向 RBM 层传播，并不断调整参数 $\theta=\{w_{ij},a_i,b_j\}$。使用传统的后向传播算法训练模型时参数收敛通常会非常缓慢，但在上述对模型参数进行微调的过程中，模型会很快收敛到一个误差极小的状态，这主要是由于上一步已对该深层架构预先进行了贪婪的预训练。

4.3 标记推断

4.3.1 转移得分

在命名实体识别任务中，捕获词语、标记与特征之间的关系往往是非常重要的。句子中的词语之间通常存在着强烈的依赖关系，句子中的标记也是以组块的形式存在。上文建立的深层架构是一个本地模型，不能捕获标记之间的依赖关系，且不能支持全局的标记序列推断。前人在运用神经网络语言模型时[12]，通常会引进一种转移得分。因此，为了对词语、标记与特征之间的关系进行建模，本文将在上文深层架构的基础上引入并计算转移得分。

给定一个输入句子 $c_{[1:n]}$，假设其标记序列为 $t_{[1:n]}$，对该标记路径条件可能性的计算需同时考虑转移得分与网络得分。转移得分 A_{ij} 表示在标记序列中从标记 i 转移到标记 j 的得分，网络得分 $f_\theta(t_i|c_{[i-2:i+2]})$ 是上文深层架构的输出，指句子中的第 i 个词语在标记 t_i 上的得分。句子 $c_{[1:n]}$ 沿着标记序列 $t_{[1:n]}$ 的总得分 s 是转移得分和网络得分之和，即

$$s(c_{[1:n]},t_{[1:n]},\theta)=\sum_{i=1}^{n}(A_{t_{i-1}t_i}+f_\theta(t_i|c_{[i-2:i+2]}))$$

给定句子 $c_{[1:n]}$ 时，可通过最大化总得分 s 发现最好的标记路径 $t_{[1:n]}$。其中，转移得分 A_{ij} 也可作为参数由后向传播算法计算得到。

4.3.2 全局推断

全局标记推断问题指给定句子 $c_{[1:n]}$ 并已知转移得分 A_{ij} 和网络得分 $f_\theta(t_i|c_{[i-2:i+2]})$ 进而估计出总得分最高的标记序列 $t^*_{[1:n]}$ 的问题，实现对句子 $c_{[1:n]}$ 中命名实体的标记。此类全局标记推断最常用的算法是维特比算法(Viterbi Algorithm)，该算法可用于寻找总得分最高的标记序列，即维特比路径。

算法 维特比算法(对给定句子 $c_{[1:n]}$ 进行全局标记推断)

输入：

给定句子：$c_{[1,n]}=(c_1,c_2,\cdots,c_n)$

转移得分：A_{ij}

网络得分：$f_\theta(t_i|c_{[i-2,i+2]})$

输出：

最优标记序列：$t^*_{[1,n]}=(t_1^*,t_2^*,\cdots,t_n^*)$

解码阶段：

初始化

$\delta_1(j)=A_{0j}+f_\theta(j|c_{[0,3]}),j=1,2,\cdots,m$

递推。对 $i=2,3,\cdots,n$

$\delta_i(l)=\max\limits_{1\leqslant j\leqslant m}\{\delta_{i-1}(j)+A_{jl}+f_\theta(l|c_{[i-2,i+2]})\}$

$l=1,2,\cdots,m$

$\Psi_i(l)=\arg\max\limits_{1\leqslant j\leqslant m}\{\delta_{i-1}(j)+A_{jl}+f_\theta(l|c_{[i-2,i+2]})\}$

$l=1,2,\cdots,m$

终止

$\max\limits_{t_{[1:n]}}(\sum\limits_{i=1}^{n}(A_{t_{i-1}t_i}+f_\theta(t_i|c_{[i-2,i+2]})))=\max\limits_{1\leqslant j\leqslant m}\delta_n(j)$

$t^*_{[1,n]}=\arg\max\limits_{1\leqslant j\leqslant m}\delta_n(j)$

返回路径

$t_i^*=\Psi_{i+1}(t_{i+1}^*),i=n-1,n-2,\cdots,1$

求得维特比路径 $t^*_{[1,n]}=(t_1^*,t_2^*,\cdots,t_n^*)$。

维特比算法结合给定句子的整个上下文进行判断，并利用递推来减少计算复杂度，从而对包含“噪音”的句子也能进行很好的解释。解码过程中，维特比算法对于给定句子中每个词语的不同标记都计算一个局部得分 $\delta_i(l)$，同时使用一个反向指针 $\Psi_i(l)$ 来指示最可能的到达该词语的某种标记路径。当完成整个计算过程后，首先在终止词语 c_n 找到最可能的标记，然后通过反向指针回溯到词语 c_1，这样回溯路径上的标记序列就是总得分最高的标记序列。

5 实验结果及分析

5.1 实验设置

本文在 Ubuntu/Linux 系统环境下用 C++和 Python 实现所有代码，完成整个模型的构建与训练。使用 Google 于 2013 年开发的一款开源工具包 word2vec 构建神经网络语言模型，该工具包的作者是 Tomas Mikolov，使用了连续的词袋模型(CBOW)[17-19]。CBOW 是一种与 NNLM 类似的模型，不同点在于 CBOW 去掉了最耗时的非线性隐层且所有词共享隐层，可无监督地训练出词特征的分布式表示和词性特征的分布式表示。使用 Python 的 Theano 库实现深度信念网络以构建出文中的深层架构。在第 1 步预训练阶段，设置训练每个 RBM 的迭代次数为 40；在第 2 步微调阶段，设置迭代次数为 20。

本文使用 PFR 人民日报标注语料库作为实验样本，该语料库是由北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作的。取实验样本的 80%作为训练语料，其余 20%作为测试语料，对样本中的人名、地名和组织机构名进行识别。在命名实体识别系统中，通常以字为单位的识别效果优于以词为单位的识别效果，且由于训练语料中不同词语的数量庞大，为降低训练词特征的分布式表示的时间，本文在对实验样本进行预处理时，使用单字切分，仅训练单字特征的分布式表示。设置 4 组对比实验如下。

实验 1　在训练语料上使用词特征和词性特征构建特征集合并训练条件随机场模型，利用得到的条件随机场模型在测试语料上进行命名实体识别，并对识别结果进行评估，记为 CRF。

实验 2　在训练语料上无监督地学习出词特征的分布式表示和词性特征的分布式表示，并利用词特征的分布式表示和词性特征的分布式表示构建并训练 4 层网络架构 375-150-50-15，在测试语料上进行命名实体识别，并对识别结果进行评估，记为 DBN4。

实验 3　在训练语料上无监督地学习出词特征的分布式表示和词性特征的分布式表示，并利用词特征的分布式表示和词性特征的分布式表示构建并训练 5 层网络架构 375-200-100-50-15，在测试语料上进行命名实体识别，并对识别结果进行评估，记为 DBN5。

实验 4　在训练语料上无监督地学习出词特征的分布式表示和词性特征的分布式表示，并利用词特征的分布式表示和词性特征的分布式表示构建并训练 6 层网络架构 375-375-200-100-50-15，在测试语料上进行命名实体识别，并对识别结果进行评估，记为 DBN6。

5.2　实验结果

本实验使用 3 个指标来衡量命名实体识别的性能：正确率、召回率和 F-值。其计算公式如下：

$$\text{正确率}(P)=\frac{\text{系统正确识别的实体个数}}{\text{系统识别的实体个数}}\times 100\%$$

$$\text{召回率}(R)=\frac{\text{系统正确识别的实体个数}}{\text{文档中的实体总数}}\times 100\%$$

$$F\text{-值}=\frac{2\times P\times R}{P+R}\times 100\%$$

对 4 组对比实验的结果进行正确率、召回率和 F-值的计算，结果如表 1 所列。

表 1　实验结果

实验	实体类型	正确率(%)	召回率(%)	F-值(%)
CRF	人名	89.58	89.86	89.72
	地名	88.14	88.61	88.37
	组织机构名	87.07	87.45	87.26
DBN4	人名	88.39	88.94	88.66
	地名	87.43	87.74	87.58
	组织机构名	86.16	86.68	86.42
DBN5	人名	89.10	89.92	89.51
	地名	88.01	88.67	88.34
	组织机构名	87.06	87.97	87.51
DBN6	人名	90.00	90.96	90.48
	地名	89.39	89.93	89.66
	组织机构名	88.30	88.93	88.61

实验 CRF 构建条件随机场模型用于命名实体识别任务，其平均正确率、召回率、F-值分别为 88.26%、88.64%、88.45%。为了减少手工特征制定的工作量，本实验仅使用词特征和词性特征构建特征集合，导致所构建的特征集合较为简单，因而造成识别效果不佳。

实验 DBN4 构建 4 层网络架构 375-150-50-15 用于命名实体识别任务，其平均正确率、召回率、F-值分别为 87.33%、87.79%、87.55%，相比于实验 CRF 分别降低了 0.93%、0.85%、0.90%。

DBN5 构建 5 层网络架构 375-200-100-50-15 用于命名实体识别任务，其平均正确率、召回率、F-值分别为 88.06%、88.85%、88.45%，相比于实验 DBN4 分别提高了 0.73%、1.06%、0.90%，且与实验 CRF 基本相当。其原因为实验 DBN4 的架构中隐层数量过少，未能在原始特征中提取出更高级的信息，实验 DBN5 通过增加隐层的数量从而提升了模型的性能。

DBN6 构建 6 层网络架构 375-375-200-100-50-15 用于命名实体识别任务，其平均正确率、召回率、F-值分别为 89.23%、89.94%、89.58%，相比于实验 CRF，分别提高了 0.97%、1.30%、1.13%。该深层架构中的第一个 RBM 使用了高斯分布对输入信息进行建模，这种方式通常会导致信息丢失，而在第一个 RBM 的隐藏层使用 375 个二值节点可最大程度地保持输入信息的完整性。

综上所述，在尽可能减少了手工特征制定工作量的情况下，使用本文构建的深层架构进行命名实体识别可达到与条件随机场模型相当的效果。

结束语　通过无监督地训练神经网络语言模型，本文得到了词特征的分布式表示和词性特征的分布式表示，并依据上述特征构建出一种深层架构用于命名实体识别任务。该方法使用了深度学习技术在原始特征上自动提取出更高级的特征，最大程度地减少了手工特征制定的工作量。在人工收集的军事语料库上进行了 4 组对比实验，结果表明该方法能够在命名实体识别任务中达到与条件随机场模型相当的效果。如何进一步减少该深层架构的训练时间，是下一步研究的重点。

参 考 文 献

[1] Tjong K, Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 142-147

[2] McCallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003: 188-191

[3] Wang Zhi-qiang. Research on Chinese named entity recognition based on conditional random fields[D]. Nanjing: Nanjing University of Science and Technology, 2006(in Chinese)
王志强. 基于条件随机域的中文命名实体识别研究[D]. 南京：南京理工大学，2006

[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507

[5] Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554

[6] Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26

[7] Hinton G E. Learning distributed representations of concepts [C]// Proceedings of the Eighth Annual Conference of the Cognitive Science Cociety. 1986, 1: 12

[8] Wang M, Manning C D. Effect of non-linear deep architecture in

sequence labeling[C]// Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP). 2013

[9] Mansur M, Pei W, Chang B. Feature-based Neural Language Model and Chinese Word Segmentation[C]// International Joint Conference on Natural Language Processing. 2013:1271-1277

[10] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research, 2003,3:1137-1155

[11] Mikolov T, Kombrink S, Burget L, et al. Extensions of recurrent neural network language model[C]// 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011:5528-5531

[12] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011,12:2493-2537

[13] Zheng X, Chen H, Xu T. Deep Learning for Chinese Word Segmentation and POS Tagging[C]// EMNLP. 2013:647-657

[14] Pei W, Ge T, Baobao C. Maxmargin tensor neural network for chinese word segmentation[C]// Proceedings of ACL. 2014

[15] Liu J S. Monte Carlo strategies in scientific computing[M]. Springer Science & Business Media, 2008

[16] Hinton G. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8):1771-1800

[17] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Neural Computation, 2014,14:1771-1800

[18] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. 2013: 3111-3119

[19] Mikolov T, Yih W, Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]// HLT-NAACL. 2013: 746-751

(上接第 196 页)

[2] Ding Z Y, Jia Y, Zhou B, et al. Survey of Influence Analysis for Social Networks[J]. Computer Science, 2014, 41(1): 48-53(in Chinese)
丁兆云,贾焰,周斌,等. 社交网络影响力研究综述[J]. 计算机科学,2014,41(1):48-53

[3] Denning P J. Computing is a natural science[J]. Communications of the ACM, 2007,50(7):13-18

[4] Jiang W Q. Zipf and the Principle of Least Effort[J]. Tongji University Journal(Social Science Section), 2005,16(1):87-95(in Chinese)
姜望琪. Zipf 与省力原则[J]. 同济大学学报(社会科学版), 2005,16(1):87-95

[5] Sun D T, He T, Zhang F H. Survey of Cold-start Problem in Collaborative Filtering Recommender System[J]. Computer and Modernization, 2012(5):59-63(in Chinese)
孙冬婷,何涛,张福海. 推荐系统中的冷启动问题研究综述[J]. 计算机与现代化,2012(5):59-63

[6] Zhang Z K, Zhou T, Zhang Y C. Tag-aware recommender systems: a state-of-the-art survey[J]. Journal of Computer Science and Technology, 2011,26(5):767-777

[7] Guan Z, Bu J, Mei Q. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009:540-547

[8] Zhang N, Zhang Y, Tang J. A tag recommendation system based on contents[C]// Proceedings of the ECML PKDD Discovery Challenge Workshop, 2009(DC09). 2009:285

[9] Xu B, Yang D, Zhang Y, et al. Relationship Bind Topic Model Toward Tag Recommendation for Micro-Blog Users[J]. Journal of Frontiers of Computer Science and Technology, 2014(3): 288-295(in Chinese)
徐彬,杨丹,张昱,等. 面向微博用户标签推荐的关系约束主题模型[J]. 计算机科学与探索,2014(3):288-295

[10] Chen Y, Lin L, Sun C J, et al. A Tag Recommendation Method for Microblog Users[J]. Intelligent Computer and Applications, 2011(5):21-26(in Chinese)
陈渊,林磊,孙承杰,等. 一种面向微博用户的标签推荐方法[J]. 智能计算机与应用,2011(5):21-26

[11] Hu J, Wang B, Liu Y. Personalized tag recommendation using social influence[J]. Journal of Computer Science and Technology, 2012,27(3):527-540

[12] Cai M S, Li X M, Yin Y T. Hybrid top-N recommendation method based on social user tag [J]. Application Research of Computers, 2013(5):1309-1311,1344(in Chinese)
蔡孟松,李学明,尹衍腾. 基于社交用户标签的混合 top-N 推荐方法[J]. 计算机应用研究,2013(5):1309-1311,1344

[13] Liao Z F, Wang C Q, Li X Q, et al. Tag Recommendation and New User Tag Recommendation Algorithms Based on Tensor Decomposition [J]. Journal of Chinese Computer Systems, 2013,34(11):2472-2476(in Chinese)
廖志芳,王超群,李小庆,等. 张量分解的标签推荐及新用户标签推荐算法[J]. 小型微型计算机系统,2013,34(11):2472-2476

[14] Wang S, Zhang L M. Mining Algorithm and Structural Analysis of Microblog Interpersonal Relationship Network Based on Tag [J]. Computer Engineering, 2014,40(5):7-11(in Chinese)
王莎,张连明. 基于标签的微博人脉网络挖掘算法和结构分析[J]. 计算机工程,2014,40(5):7-11

[15] Atallah M J. Algorithms and Theory of Computation Handbook [M]. CRC Press, 1998

[16] Chen L F, Mark-Liao H Y, Ko M T, et al. A new LDA-based face recognition system which can solve the small sample size problem[J]. Pattern Recognition, 2000,33:1713-1726

[17] Heinrich G. Parameter estimation for text analysis[R]. 2004

[18] Liu Yang, Qiu Ming-hui, Gottipati S, et al. CQARank: Jointly Model Topics and Expertise in Community Question Answering [C]// Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013). 2013:99-108

[19] 定型数据集和测试数据集[OL]. http://msdn.microsoft.com/zh-cn/library/bb895173.aspx