基于 SDAs 的人物关系抽取方法研究

珠 杰 洪军建
(西藏大学计算机科学系 拉萨 850000)

摘 要 针对人物关系语料缺乏的问题,研究了基于互动百科的自动标注方法;针对传统浅层机器学习模型特征表示能力差的问题,提出了基于深度神经网络模型 SDAs 的人物关系抽取方法。重点研究了多个特征组合的人物关系抽取效果以及不同深度 SDAs 网络的人物关系抽取效果。根据实验分析,F 系数可达到 73.75%。

关键词 社会网络,人物关系抽取,降噪自动编码器,深度学习
中图法分类号 TP393 **文献标识码** A

Research on Method of Personal Relation Extraction under SDAs

ZHU Jie HONG Jun-jian
(Department of Computer Science, Tibet University, Lasa 850000, China)

Abstract For the lack of corpus issue in personal relation, this paper studied the methods of automatic tagging based on HUDONG pedia; for poor ability to express feature issue in shallow machine learning models, we proposed the method of personal relation extraction under deep learning model SDAs and focused on the effect of personal relation extraction with combination features and effect of personal relation extraction with different depths in SDAs network. F factor can reach 73.75% through experiment analysis.

Keywords Social network, Extraction of personal relation, SDAs, Deep learning

人物关系抽取是信息抽取领域的热门研究方向之一[1],指的是从非结构化或半结构化的文本中识别用户感兴趣的人物关系,并以结构化的形式进行存储的过程。随着计算机技术的不断发展,互联网在社会生活的各个方面发挥着越来越重要的作用,是一个信息量极其丰富的知识资源宝库[2]。其中蕴含的人物及其关系信息在政治、经济、社会、军事等各领域具有重要的应用价值,人们早已开始广泛关注此问题并纷纷展开相关研究。人物关系抽取指的是从纯文本中发现人名实体对之间存在的语义关系,是社会网络构建与分析等一系列研究的基础工作。本文以互联网的 Web 资源为研究对象,1)借助互动百科等研究多方法协同标注人物关系语料的方法;2)利用深度学习模型 SDAs 研究多个特征下的人物关系抽取方法。

本文第 1 节介绍了人物关系抽取的相关方法;第 2 节研究了多个方法协同的人物关系语料自动生成和标注方法;第 3 节研究了多种特征组合下深度学习人物关系的抽取方法;第 4 节是特征组合和不同深度层次的实验及分析结果;最后是结论与展望。

1 相关工作

当前,信息抽取的研究方法众多,主流方法是基于机器学习的方法,本文亦采用此方法进行人物关系抽取研究。目前,基于机器学习的方法中,比较经典的方法有两种:基于特征的方法、基于核函数的方法。下面分别介绍这两种方法及社会网络构建方法的相关研究。

(1)基于特征的方法。基于特征向量抽取[3]的方法主要从关系实例的实体上下文、词性、句法等信息中抽取一系列特征,并训练一个分类器(朴素贝叶斯、支撑向量机、最大熵等),从而完成关系抽取任务。Kambhatla 等[4]从文本中选取了实体的上下文、句法结构、语义等多种特征,利用最大熵模型对人物关系的抽取效果进行了研究,在自动内容抽取(ACE)的评测中取得了比较有竞争力的结果。Jiang 等[5]给出了基于图表示法的特征空间的一般定义,并在这个框架下对关系实例的 3 种不同表述形式和不同复杂度的特征进行了探讨,实验结果表明,相较于单一特征,复合特征可以在一定程度上提升关系抽取的准确率。董静等[6]根据实体关系将其分为包含实体关系以及非包含实体关系的方式,并对这两种关系采用不同的句法特征集,利用条件随机场模型(CRF)进行了关系抽取研究,在 ACE2007 语料库上的实验结果表明,该划分方法有效提高了实体关系抽取的性能。

(2)基于核函数的方法。基于特征的方法需要考虑如何将非线性结构(句法)转换成线性结构,而基于核函数的方法[7]可以利用核函数直接计算两个非线性结构的相似度,不需要抽取特征,从而也没有维数灾难问题。Bunescu 等[8]提出,两个命名实体之间的关系所需的信息一般包含在依存图中两个实体之间的最短路径上,通过在 ACE 新闻语料上的大类关系实验表明,最短依存路径核优于依存树核。Zhang M 等[9]利用复合核函数对基于两个实体的关系抽取进行了

本文受国家自然科学基金项目(61262058),西藏大学高原学者-珠杰项目资助。
珠 杰(1973-),男,副教授,硕士生导师,主要研究方向为藏文信息处理与数据挖掘等,E-mail:rocky_tibet@qq.com。

研究,在 ACE 语料上的实验表明,该方法优于先前报道的最好方法。Huang 等[10]基于 ACE 2007 语料库对卷积树核和最短依存树核在中文关系抽取中的效果进行了探索,实验结果表明基于核的方法对中文关系抽取来说是一种有效的方法。

(3)基于 Web 网页的社会网络构建。Kautz 等[11]基于 Web 构建了一种名为 Referral Web 的社会网络系统,该系统的目标是通过对 Web 上的数据进行挖掘来构建一个社会网络模型,并利用该模型开发相关的工具来帮助用户快速地对所需的人物信息进行定位。Mika 等[12]开发了一个用于在线社会网络的 Flink 系统,该系统集抽取、聚合和可视化于一体,通过将网页、电子邮件、出版档案、用户创建的个人信息文件(FOAF)等作为语料,提出了一种基于语义 Web 的社会网络分析的新方法。Chang 等[13]提出了基于概率话题模型的实体描述信息和实体间关系描述信息的推断方法,在 Wikipedia 语料上的实验结果表明,利用该模型可以构造带有标注信息的人物关系图,并且可以对人物关系进行有效推断。姚从磊等[14]定义了 6 种人物关系(Wife, Husband, Father, Mother, Children 和 Close Friend 等),并采用模拟退火算法发掘网页中蕴涵的人物社会关系的最小描述模式集合,进而利用 Web 信息的冗余性提取人物关系信息,其人物社会关系虽然丰富,但同时会带来数据稀疏问题。

(4)基于纯文本的社会网络构建。利用纯文本进行社会网络构建,不仅可以获取到丰富可靠的人物关系,而且可以结合指代消解的相关技术解决人名歧义问题。Jing 等人[15]通过命名实体识别、关系抽取、事件抽取、指代消解等技术从特定领域的口语转录文本中提取相关的人物关系和事件,并构建相应的社会网络。实验结果表明,由于语料本身质量较差,构建的社会网络性能 F 值只达到了 30% 左右。Elson 等[16]对从文学作品中直接抽取社会网络的方法进行了研究,通过角色名称识别和对话检测的方式来确定一段对话中的两个角色,并在此基础上确定两者之间的人物关系,构造的社会网络性能 F 值为 47%。不过,该方法只适用于从小说文本的角色对话中抽取人物关系。Camp 等[17]从 574 篇历史自传中抽取社会活动家之间的关系,将这种关系按情感极性分为正面、中性和反面 3 种,除了考虑自传中人物的共现特征外,还考虑了部分词汇特征,利用 SVM 分类器对人物关系进行分类,并构建了社会网络。

根据以上人物关系抽取方法的对比分析,基于特征分析方法需要构造大量的特征,并且存在特征稀疏的问题;基于核函数的方法需要花费大量的时间来计算树核之间的对比关系。鉴于两种机器学习方法存在的不足,本文根据深度学习的端到端学习特征的特点,利用经典模型堆叠降噪自动编码器[18](SDAs, Stacked Denoising Autoencoders)来进行人物关系抽取研究,探讨深度网络对复杂问题的解决能力。

2 语料自动生成及标注方法

近些年来,研究关系抽取所需的语料主要来源于维基百科、ACE 语料库和一些领域自制的语料库(如生物领域)等。研究人员利用这些语料库进行关系抽取研究,并取得了可喜的研究成果。目前,针对人物关系抽取的语料还很少见,研究人员不得不人工标注人物关系数据集来满足研究所需的语料要求。当然,采用人工标注的方式不仅需要耗费大量的人力和物力,而且很难完成大规模语料标注的任务。因此,本节利用大量的互联网数据及相关百科知识库,并通过多种方法协同来完成人物关系语料的自动生成与标注。

首先,设计了数据抓取模块、网页正文抽取模块、人物关系语料生成模块和人物关系语料自动标注模块;其次,利用两种语言处理平台和互动百科协同来实现人物关系语料的自动生成和标注。具体过程描述如下:

(1)数据抓取模块

本文的网络爬虫利用开源爬虫软件 Nutch 在单台机器上进行,通过使用多线程技术提升爬虫程序的网页抓取效率。

(2)网页正文抽取模块

本文采用哈尔滨工业大学信息检索研究中心陈鑫提出的基于行块分布函数的通用网页正文抽取算法来进行网页正文抽取。

(3)人物关系语料生成模块

通过对网页进行正文抽取,获得了大量的初级文本语料。这些语料是以平常所见到的文本形式进行存储的,包含文档标题和正文的各个段落。此种形式的语料是无法满足后续处理要求的。因此,需要对这些抽取到的正文文本做断句、分词、词性标注和人名实体识别处理。在此,采用 NLPIR 汉语分词系统和语言技术平台(LTP)对每篇文档都各做一次处理,分别获取到相应的断句、分词、词性标注及人名实体识别结果。通过协同 NLPIR 和 LTP 这两个平台的人名实体识别结果,获取到更加准确的人名实体。经过预处理之后,语料集合可表示为式(1)和式(2)的形式:

$$SSA=\{S_{1a},S_{2a},\cdots,S_{ia},\cdots,S_{na}\} \tag{1}$$

$$SSB=\{S_{1b},S_{2b},\cdots,S_{ib},\cdots,S_{nb}\} \tag{2}$$

其中,SSA 指的是 NLPIR 汉语分词系统对整个初始语料处理的结果,SSB 指的是 LTP 平台对整个初始语料处理的结果。式中每一个元素可表示为式(3)和式(4)所示的形式:

$$S_{ia}=(ws_{1a}/pos_{1a},ws_{2a}/pos_{2a},\cdots,ws_{ka}/pos_{ka},pe_1,pe_2,\cdots,pe_q) \tag{3}$$

$$S_{jb}=(ws_{1b}/pos_{1b},ws_{2b}/pos_{2b},\cdots,ws_{ta}/pos_{ta},pe_1,pe_2,\cdots,pe_r) \tag{4}$$

其中,ws 表示分词得到的词语,pos 表示词性,pe 表示人名。

为了获得更准确的人名实体识别结果,将两个平台与互动百科结合在一起实现人物关系语料的自动生成。算法的基本思想如下:若两个平台的人名实体识别结果一致,则认为人名识别正确;否则将该人名提交互动百科,若互动百科查询不到该人名信息或该实体的性别、民族等信息,则认为其不是一个人名实体,在人名列表中删除该实体。为了简化人名实体之间的关系,只保留含有两个人名的句子。过程如下:判断一条语料中的人名实体数量,若大于 2 或者小于 2,则该条语料舍弃;否则,保留该条语料。人物关系语料自动生成算法如图 1 所示。

经统计,SSA 包含语料 356892 条,SSB 包含语料 328436 条,经过该算法处理后得到包含两个人名实体的语料 173827 条。

```
输入：SSA，SSB
输出：人名实体对识别结果语料
1. for i from 1 to n //人名实体矫正算法
2.    for each pe_q in S_ia(S_ia∈SSA)
   //(for each pe_r in S_ib，to do this operation)
3.       if S_ib(S_ib∈SSB) not contains pe_q
4.          then pe_q 提交互动百科
5.             if pe_q not in 互动百科 or not pe_q person's information
6.                from S_ia delete pe_q
7. for i from 1 to n //人名实体对获取过程
8.       if S_ia person number !=2//(S_ib to do this operation)
9.       delete S_ia from SSA//or delete S_ib from SSB
10.  else
11.      SS+=SSA//SS+=SSB
```

图 1　人物关系语料自动生成算法

(4)人物关系语料标注模块

经过上面的处理，可以得到如式(5)形式的语料集：

$$SS=\{S_1,S_2,S_3,\cdots,S_m\} \tag{5}$$

其中，

$$S_i=(ws_1/pos_1,ws_2/pos_2,\cdots,ws_k/pos_k,pe_1,pe_2) \tag{6}$$

本文定义了 3 种人物关系类型，分别为家庭关系、社会关系和其他关系，并设计了每个关系类型的知识库。在标注模块中，提出了基于互动百科的人物关系类型自动标注算法，算法通过抽取互动百科的人物关系类型来实现对人物关系语料的自动标注，具体算法如图 2 所示。

```
输入：人名实体对算法的识别结果
输出：标注关系类型的实例
1. for each S_i in SS
2.    Flag =FALSE
3.    提交 pe_1 给互动百科//(pe_1∈S_i)
4.    返回结果：〈pe_1，pe_12，r_12〉，…，〈pe_1，pe_1n，r_1n〉
5.    for k from 1 to n
6.       if pe_1k equals pe_2//(pe_2∈S_i)
7.          then 得到〈pe_1，pe_2，r_1k〉
8.          Flag=TRUE
9.    end for
10.   if Flag==FALSE
11.      提交 pe_2 给互动百科
12.      返回结果：〈pe_2，pe_22，r_22〉，…，〈pe_2，pe_2m，r_2m〉
13.      for k from 1 to m
14.         if pe_2k equals pe_1
15.            then 得到〈pe_1，pe_2，r_2k〉
16.            Flag=TRUE
17.      end for
18.   if Flag==TURE
19.      S_i=(ws_1/pos_1，…，ws_k/pos_k，pe_1，pe_2，r_1k/r_2k)
```

图 2　人物关系类别自动标注算法

该算法首先将一个人名实体 pe_1 提交给互动百科，抽取其中的人物关系信息，如果得到了关系三元组〈pe_1，pe_2，r_{1k}/r_{2k}〉，则用 r_{1k}/r_{2k} 标记语料，从而得到如下的语料表示形式 S_r：

$$S_r=(ws_1/pos_1,ws_2/pos_2,\cdots,ws_k/pos_k,pe_1,pe_2,r_{1k}/r_{2k}) \tag{7}$$

如此迭代整个 SS 语料集，便可完成对整个语料集的关系类型标注，得到的标注语料集可表示为 SS_r：

$$SS_r=\{S_{r1},S_{r2},S_{r3},\cdots,S_{rm}\} \tag{8}$$

通过对 173827 条只包含两个人名实体的语料执行该算法，得到可用于人物关系标注的语料 27891 条。

3　基于 SDAs 的人物关系抽取方法

堆叠降噪自动编码器(Stacked Denoising Autoencoders，SDAs)是一种深度神经网络模型，与传统浅层机器学习模型的主要不同之处在于可以对特征进行分层表示，以便获得不同抽象程度的特征。本文主要通过堆叠降噪自动编码器研究了基于单句的人物关系抽取效果，并通过实验探索了堆叠降噪自动编码器的不同网络参数对人物关系抽取评估指标的影响。

3.1　特征抽取

本文主要利用了实体上下文的词特征、词性特征、实体之间的相对位置特征、依存句法特征[19]、语义特征[20]进行特征向量构造，并探究这些特征对人物关系抽取效果的贡献。本研究的语料来源于基于互动百科的人物关系语料自动生成与标注算法获取到的人物关系语料，第 i 条语料的形式如式(9)所示：

$$S_r(i)=(ws_1/pos_1,ws_2/pos_2,\cdots,ws_k/pos_k,pe_1,pe_2,r_t) \tag{9}$$

该语料由 4 个部分组成，分别是分词得到的词语 ws、词性标注 pos、人名实体 pe_1 与 pe_2 以及关系类型 r_t。

(1)词特征：词特征指的是人名实体前后出现的词语，即实体的上下文。提取人名实体前后出现的若干词语，窗口大小用 win 表示。根据提取到的这些词语，可以构造形如式(10)所示的词典：

$$D=(d_1,d_2,d_3,\cdots,d_i,\cdots,d_n) \tag{10}$$

其中，d_i 表示词典中的第 i 词语，$1\leqslant i\leqslant n$，n 是词典的规模大小，即词典中包含的总词数。那么，每一条训练语料的特征向量可表示为：

$$V(S_r(i))=(v_{i1},v_{i2},v_{i3},\cdots,v_{in}) \tag{11}$$

其中，$S_r(i)$表示第 i 条训练语料。词典中各项的权重设定采用布尔权重，那么，词特征 ws_k 的值满足式(12)所示的关系：

$$v_{ik}=\begin{cases}1, & ws_k\in S_r(i)\\ 0, & ws_k\notin S_r(i)\end{cases} \tag{12}$$

经过词特征的处理，词特征就变成了高维的稀疏空间。为了缓解句子中词特征过少导致的数据高度稀疏问题，利用同义词词林对句子中的特征词进行同义词扩展。

(2)词性特征：词性是词汇的基本语法属性，通常也被称为词类。从语料 S_r 中提取到词特征的词性表示成式(13)形式：

$$FP(S_r)=(pos_1,pos_2,\cdots,pos_k) \tag{13}$$

在此，利用词性特征构造特征向量，将词性符号量化为便于计算的数值。在 NLPIR 汉语分词系统中共定义了 22 种词性，本文对这 22 种词性从 0 至 21 依次编码。窗口 win 中词特征词对应的词性特征可以表示为一个 22 维的词性向量，如式(14)所示：

$$V(pos_i)=(p_{i1},p_{i2},\cdots,p_{ik}) \tag{14}$$

其中，$k=22$，该向量是一个布尔向量，第 i 条语料 $S_r(i)$中词特征的第 k 个词性特征项 pos_k 的权重取值规则如式(15)所示：

$$p_{ik}=\begin{cases}1, & pos_k\in c_i\\ 0, & pos_k\notin c_i\end{cases} \tag{15}$$

(3)实体之间的相对位置特征:实体之间的相对位置特征对关系抽取系统的精度有很大的影响,是进行关系抽取时应当考虑的一类非常重要的特征信息。实体之间的相对位置特征可以表示为一个 2 维的特征向量,如式(16)所示:

$$V(loc)=(loc_1,loc_2) \tag{16}$$

其中,loc_1 表示人名实体 P_1 和 P_2 之间相邻的相对位置关系,loc_2 表示人名实体 P_1 和 P_2 之间分离的相对位置关系。若语料中两个人名实体的相对位置关系为相邻,则 $loc_1=1$,$loc_2=0$,即 $V(loc)=(1,0)$;反之,如果为分离关系,则 $loc_1=0$,$loc_2=1$,即 $V(loc)=(0,1)$。

(4)依存句法特征:本文采用哈尔滨工业大学的语言技术平台(LTP)对语料做进一步的句法分析,以验证依存句法特征对人物关系抽取性能的影响。LTP 平台共定义了 14 种依存关系类型,依存关系类型的编码为 0 至 13。由于不同句子的依存路径长度不一样,将导致所构造的依存句法向量的维度有差异。为了得到维度一致的依存句法向量,以维度最大的依存句法特征向量为范本,对其他依存句法向量的规模进行扩展,向量的扩展部分统一取 14 作为其编码。依存句法特征的向量可形式化地表述为式(17):

$$V(dp_i)=(dp_{i1},dp_{i2},dp_{i3},\cdots,dp_{ik}) \tag{17}$$

(5)语义特征:语义角色标注(Semantic Role Labeling,SRL)是一种浅层的语义分析技术,标注句子中某些短语为给定谓词的论元(语义角色),如施事、受事时间和地点等。LTP 共定义了 21 种语义角色,因此语义特征可以用一个 21 维的向量进行表示,如式(18)所示:

$$V(srl_i)=(s_{i1},s_{i2},\cdots,s_{ik}) \tag{18}$$

其中,srl_i 表示第 i 个语句的语义角色信息,$k=21$,向量中表示相应语义角色的向量元素的值设为 1,其他元素置为 0。

3.2 SDAs 的人物关系抽取

SDAs 网络是一种无监督机器学习模型,是深度学习领域的经典模型。本文首先研究了单项特征进行组合的多特征的人物关系识别效果;其次研究了多特征条件下,SDAs 网络的隐含层数对人物关系识别的效果。

(1)网络的预训练

深度网络的预训练是一个逐层贪婪寻优的过程,是一个无监督的学习过程。无监督的预训练可以充分利用大量的无标签语料将网络的参数快速定位到接近全局最优的范围,相比于参数随机初始化的 ANN 网络,SDAs 可以避免陷入局部最优。在预训练阶段,网络接收上面构造的特征向量,通过多次迭代的形式进行参数优化。上节收集的 356892 条语料中,标注的语料 27891 条,未标注的语料 329001 条。未标注部分作为预训练的语料,用于训练 SDAs 网络。

(2)网络的微调

通过大量的无标签数据将网络参数快速定位到接近全局最优的范围之后,再利用有标签数据对网络的参数进行微调,这样可以显著提升网络的精度。本文通过人物关系语料自动生成系统获取到的 27891 条标签数据对 SDAs 网路进行微调。输入数据是特征向量,目标值是该特征向量的标签值,即本文自定义关系类型:家庭关系(HC)、社会关系(SC)和其他关系(OC)。为了方便计算,特对标签进行了数值化处理:数值 0 代表家庭关系、数值 1 代表社会关系、数值 2 代表其他关系。通过预训练,已将网络参数定位到了接近全局最优的范围,此时再进行微调,可以很快地使网络的参数达到全局最优。该过程采用随机梯度下降算法,而利用分类器的输出与目标值的差异求解损失函数,然后利用梯度下降来调整参数。

(3)人物关系抽取

基于 SDAs 网络进行人物关系抽取的具体过程如下:

1)对 SDAs 进行无监督预训练,将网络的参数快速定位到接近全局最优的范围。

2)利用有标签数据对网络的参数进行微调,以提升网络的精度。

3)对于给定的一个输入语句,首先进行分词、词性标注以及人名识别之后,将输入表示成式(19)形式:

$$S_i=(ws_1/pos_1,ws_2/pos_2,\cdots,ws_k/pos_k,pe_1,pe_2) \tag{19}$$

其中,ws_i 表示第 i 个词语,pos_i 表示第 i 个词语的词性,pe_1 和 pe_2 是两个人名实体。如果无法表示成该形式,则将不对该输入进行人物关系抽取操作,并返回一个空的关系三元组;否则,在此基础上,根据上节的特征向量构造方法构造式(20)所示形式的特征向量:

$$V=(v_1,v_2,\cdots,v_n) \tag{20}$$

然后把构造的向量输入 SDAs 网络。

4)SDAs 网络对输入的向量进行降维处理,得到对输入特征的压缩表示。

5)把 SDAs 的输出作为 softmax 分类器的输入,经过 softmax 的分析,预测得到相应的关系类型。

6)将得到的关系三元组在文件中存储,以便作为构建社会网络的数据。

4 实验结果与分析

本实验的数据语料是通过本文设计的人物关系语料自动生成系统在互联网新闻站点获取到的。经统计,在获取到的 356892 条语料中,标注的语料 27891 条,未标注的语料 329001 条。用未标注语料对 SDAs 网络进行预训练,用标注语料对网络进行微调。实验按照 4∶1 的原则,将标注语料划分为训练集和测试集,即训练集包含语料 22313 条,测试集包含语料 5578 条。在实验中,窗口 win 设为 2,表示人名实体前后分别获取 2 个词特征,词性特征获取也是词特征对应的词性。人物关系抽取的评价指标借鉴了信息检索领域的准确率(Precision)、召回率(Recall)、F 指数(F-Measure)。下面按照特征组合实验、网络深度实验对实验结果进行分析。

4.1 特征组合实验

本组实验将在词特征的基础上依次组合词性特征(pos+)、位置特征(loc+)、依存句法特征(dp+)和语义特征(srl+)。例如 srl+表示 ws+pos+loc+dp+srl 的特征的组合,dp+表示 ws+pos+loc+dp 的特征组合。使用 1 层网络进行测试,表 1 列出了特征组合之后人物关系识别的效果。

表 1 组合特征的人物关系识别结果/%

特征	评估指标		
	P	R	F
ws+	55.04	45.72	49.92
pos+	58.18	44.92	50.70
loc+	72.34	57.29	63.94
dp+	75.64	58.73	66.12
srl+	73.90	58.19	65.11

表 1 表明，在词特征的基础上添加相应的词性特征，对关系识别性能的影响不是很大，其中关系识别的准确率提升了 3 个百分点左右，而召回率下降了 0.8 个百分点，F 值提升了 0.78 个百分点。其原因可能是词性特征和词特征存在较强的相关性，词性特征所带来的新信息比较少。然后添加了相对位置特征，关系识别的准确度提升了 14 个百分点左右，召回率提升了 13 个百分点左右，F 值提升了 13 个百分点左右，这表明人物实体的相对位置特征对人物关系识别的影响是显著的，是对系统性能提升起着关键作用的一类特征，特别是对系统召回率的提升有着十分明显的效果。然后添加了依存句法特征，准确率提升了 3 个百分点以上，召回率提升了 1.5 个百分点左右，F 值提高了 2 个百分点以上，有一定的提升效果但不是很明显，原因可能在于：1）目前句法分析技术不是很成熟，所获取的句法特征不能很好地描述语句的句法关系；2）与上述的特征存在相关性，带来的新信息量不多。最后添加了语义特征，准确率下降了 1.7 个百分点左右，召回率下降了 0.5 个百分点左右，F 值下降了 1 个百分点左右，也就是说语义特征带来了系统性能的整体下滑，究其原因，可能在于：1）语义分析技术处于起步阶段，为系统引入了较多的错误信息；2）语义分析技术是基于语法分析技术的，语法分析的错误信息被再一次引入系统，导致人物关系识别性能下降。

4.2　网络深度实验

理论研究表明，深层结构比浅层结构的学习能力更强大，能够学习出高度抽象的特征，有助于复杂的机器学习问题的求解[11]。基于此，本实验研究了 SDAs 网络的深度对人物关系识别效果的影响。表 2 列出了 SDAs 网络的深度分别为一层、二层、三层、四层的实验结果。

表 2　不同网络深度的 SDAs 人物关系识别效果/%

特征	一层	二层	三层	四层
	P/R/F	P/R/F	P/R/F	P/R/F
ws+	55.04/45.72/49.92	57.92/46.11/51.34	58.28/45.33/51.00	58.17/46.35/51.59
pos+	58.18/44.92/50.70	58.76/47.39/52.47	61.94/47.55/53.80	62.72/46.29/53.27
loc+	72.34/57.29/63.94	73.17/61.64/66.91	72.83/63.82/68.03	76.81/63.22/69.40
dp+	75.64/58.73/66.12	77.18/64.91/70.51	79.36/67.19/72.77	81.94/64.24/72.02
srl+	73.90/58.19/65.11	73.28/71.47/72.36	75.66/69.52/72.46	74.95/72.59/73.75

由表 2 可知，词特征在网络层数为三层时取得最大准确率 58.28%，网络层数为四层时取得最大召回率 46.35%，综合来看，网络层数为四层时 F 值取得最大值 51.59%；在词特征的基础上添加词性特征，网络层数为四层时取得最大准确率 62.72%，网络层数为三层时取得最大召回率 47.55%，综合来看，网络层数为三层时取得最大 F 值 53.80%；添加相对位置特征后，网络层数为四层时取得最大准确率 76.81%，网络层数为三层时取得最大召回率 63.83%，综合来看，网络层数为四层时取得最大 F 值 69.4%；添加依存句法特征后，网络层数为四层时取得最大准确率 81.94%，网络层数为三层时取得最大召回率 67.19%，综合来看，网络层数为三层时取得最大 F 值 72.77%；最后，添加语义特征后，网络层数为三层时取得最大准确率 75.66%，网络层数为四层时取得最大召回率 72.59%，综合来看网络层数为四层时取得最大 F 值 73.75%。由以上分析不难看出，随着特征逐渐复杂化，最大准确率、最大召回率、最大 F 值始终在网络深度为三层和四层时徘徊，相比于网络深度为一层和二层时，结果有明显的提高。对语义特征而言，在网络深度为一层、二层和三层时，该特征导致准确率、召回率以及 F 值整体下降，但是当网络层数为四层时，却使得召回率和 F 值得到了提高，这在一定程度上表明，增加网络深度可以提高网络对复杂问题的解决能力，随着问题复杂度的增加，浅层网络表现出了一定的局限性。

结束语　人物关系抽取作为一种收集人际关系的手段，与人们的日常生活和工作学习息息相关，具有广泛的应用价值。本文研究了多方法协同的人物关系语料自动生成和标注方法，并利用 SDAs 模型研究了人物关系抽取方法，虽然取得了一些成果，但还是存在一定的不足，需要进一步的研究与完善。下一步的工作可以从以下几个方面展开：

（1）在特征选择上，本文只是选取了词特征、词性特征、实体相对位置特征、依存句法特征以及语义特征 5 个特征，下一步可以考虑寻找更多的特征来提升人物关系抽取的效果；

（2）实验中发现，单机环境下 SDAs 网络的训练速度是非常缓慢的，下一步可以考虑设计基于 MapReduce 架构的并行 SDAs 网络，以提高网络的训练速度，节省时间。

参 考 文 献

[1] 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 on Knowledge Discovery and Data Mining. 2014: 601-610.

[2] RIEDEL S, YAO L, MCCALLUM A, et al. Relation extraction with matrix factorization and universal schemas [C]// Proceedings of NAACL-HLT 2013. Atlanta, Georgia, 2013: 74-84.

[3] ZHAO G, WU Y, CHEN F, et al. Effective feature selection using feature vector graph for classification [J]. Neurocomputing, 2015, 151: 376-389.

[4] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C]// Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. USA, 2004: 178-181.

[5] JIANG J, ZHAI C X. A Systematic Exploration of the Feature Space for Relation Extraction [C]// Proceedings of NAACL HLT 2007. Rochester, NY, 2007: 113-120.

[6] 董静，孙乐，冯元勇，等. 中文实体关系抽取中的特征选择研究 [J]. 中文信息学报，2007，21(4)：80-85.

[7] DING S, ZHANG Y, XU X, et al. A novel extreme learning machine based on hybrid kernel function [J]. Journal of Computers, 2013, 8(8): 2110-2117.

[8] BUNESCU R C, MOONEY R J. A shortest path dependency kernel for relation extraction [C]// Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, BC, 2005: 724-731.

（下转第 150 页）

由上述实验可知，准确率高的 CNN 按照图像属性分级学习特征，我们可以利用可视化工具寻找准确率高且结构小的网络，以保证训练的效率。

将这个结论推广到其它图像集，如要进行人脸识别时，则可以根据人脸图像的结构层次设计我们想要的准确率高且结构小的网络。

结束语 本文的贡献主要在：首先对国内外 CNN 反卷积可视化研究进行了总结梳理，其次利用数值求解改进了经典反卷积可视化模型，并构造了具有不同结构层次的数据集来探究准确率高的 CNN 与分级学习的关系。实验结果表明，准确率高的网络按照分级提取特征。针对这种结果，可以进一步在更大的数据库上进行试验探究来验证结果，并利用这种结果，进一步探究准确率高与网络结构之间存在的关系。

参 考 文 献

[1] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[2] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436-444.

[3] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012：1097-1105.

[4] 王晓刚. 深度学习在图像识别中的研究进展与展望[OL]. http://chuangsong.me/n/2103247.

[5] OUYANG W L，WANGX G. Joint deep learning for pedestrian detection[C]// ICCV. New York：IEEE，2013：2056-2063.

[6] CHEN Y N，HAN C C，WANG C T，et al. The application of a convolution neural network on face and license plate detection [C]// The 18th International Conference on Pattern Recognition. Hongkong：IEEE，2006，4：552-555.

[7] NGUYEN A，YOSINSKI J，CLUNE J. Deep neural networks are easily fooled：High confidence predictions for unrecognizable images[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2015：427-436.

[8] FAWZI A，FAWZI O，FROSSARD P. Analysis of classifiers' robustness to adversarial perturbations[J/OL]. Available：http://arxiv.org/abs/1502.02590.

[9] SCHWENK H，BOUGARER F，BARRAULT L. Efficient training strategies for deep neural network language models[C]// NIPS workshop on Deep Learning and Representation Learning. 2014.

[10] ZEILER M D，KRISHNAN D，TAYLOR G W，et al. Deconvolutional networks[C]// 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2010：2528-2535.

[11] KAVUKCUOGLU K，SERMANET P，BOUREAU Y，et al. Learning onvolutional feature hierarchies for visual recognition[C]// Proceedings of Advances in Neural Information Proces-sing Systems，2010. New York：Curran Associates，Inc.，2010：1090-1098.

[12] ZEILER M D，TAYLOR G W，FERGUS R. Adaptive deconvolutional networks for mid and high level feature learning[C]// 2011 IEEE International Conference on Computer Vision (ICCV). IEEE，2011：2018-2025.

[13] DOSOVITSKIY A，BROX T. Inverting Visual Representations with Convolutional Networks[J]. arXiv preprint arXiv：1506.02753，2015.

[14] ZEILER M D，FERGUS R. Visualizing and understanding convolutional networks[M]// Computer Vision-ECCV 2014. Springer International Publishing，2014：818-833.

[15] YOSINSKI J，CLUNE J，NGUYEN A，et al. Understanding neural networks through deep visualization[J]. arXiv preprint arXiv：1506.06579，2015.

[16] BOUREAU Y L，PONCE J，LECUN Y. A Theoretical Analysis of Feature Pooling in Visual Recognition[C]// International Conference on Machine Learning. 2010：459-459.

[17] FARABET C，COUPRIE C，NAJMAN L，et al. Learning hierarchical features for scene labeling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35(8)：1915-1929.

(上接第 145 页)

[9] ZHANG M，ZHANG J，SU J，et al. A composite kernel to extract relations between entities with both flat and structured features[C]// Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney，2006：825-832.

[10] HUANG R，SUN L，FENG Y. Study of kernel-based methods for Chinese relation extraction[M]// Lecture Notes in Computer Science：Information Retrieval Technology. 2008：598-604.

[11] BENGIO Y. Learning deep architectures for AI[J]. Foundations and trends® in Machine Learning，2009，2(1)：1-127.

[12] KAUTZ H，SELMAN B，SHAH M. The hidden Web[J]. AI magazine，1997，18(2)：27.

[13] MIKA P. Flink：Semantic web technology for the extraction and analysis of social networks[J]. Web Semantics：Science，Services and Agents on the World Wide Web，2005，3(2)：211-223.

[14] CHANG J，BOYD-GRABER J，BLEI D M. Connections between the lines：augmenting social networks with text[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris，France，2009：169-178.

[15] 姚从磊，邸楠. 一种基于 Web 的大规模人物社会关系提取方法[J]. 模式识别与人工智能，2007，20(6)：740-744.

[16] JING H，KAMBHATLA N，ROUKOS S. Extracting social networks and biographical facts from conversational speech transcripts[C]// Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague，2007：1040-1047.

[17] ELSON D K，DAMES N，MCKEOWN K R. Extracting social networks from literary fiction[C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. PA，USA，2010：138-147.

[18] VAN DE CAMP M，VAN DEN B A. A link to the past：constructing historical social networks[C]// Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Portland，Oregon，2011：61-69.

[19] 郭振，张玉洁，苏晨，等. 基于字符的中文分词，词性标注和依存句法分析联合模型[J]. 中文信息学报，2014，28(6)：1-8.

[20] 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.