·开发研究与工程应用· 文章编号：1000-3428(2018)06-0305-06 文献标志码：A 中图分类号：TP391

基于堆栈降噪自编码的维吾尔语事件共指关系识别

王淑媛[a]，田生伟[a]，禹 龙[b]，冯冠军[c]，艾山·吾买尔[d]，李 圃[e]，赵建国[c]

（新疆大学 a. 软件学院；b. 网络中心；c. 人文学院；d. 信息科学与工程学院；e. 语言学院，乌鲁木齐 830046）

摘 要： 结合维吾尔语的语言特点，基于堆栈降噪自编码（SDAE），提出一种新的维吾尔语事件共指关系识别方法。将维吾尔语事件两两构成候选事件对，抽取事件基本属性、触发词、事件距离等 9 项特征，利用 Word Embedding 富含语义信息的特性，将其计算得到的维吾尔语事件触发词语义相似度作为特征之一，训练 SDAE 模型，将 SDAE 的输出作为 softmax 层的输入，从而分类完成维吾尔语事件共指关系识别任务。实验结果表明，与浅层机器学习模型支持向量机相比，基于深度学习机制的 SDAE 模型更适用于维吾尔语事件共指关系识别任务，并提升了识别性能。

关键词： 共指关系；维吾尔语；语义相似度；堆栈降噪自编码；深度学习

中文引用格式： 王淑媛，田生伟，禹 龙，等. 基于堆栈降噪自编码的维吾尔语事件共指关系识别[J]. 计算机工程，2018，44(6)：305-310.

英文引用格式： WANG Shuyuan，TIAN Shengwei，YU Long，et al. Identification of Uyghur event coreference relationship based on stacked denoising autoencoder[J]. Computer Engineering，2018，44(6)：305-310.

Identification of Uyghur Event Coreference Relationship Based on Stacked Denoising Autoencoder

WANG Shuyuan[a]，TIAN Shengwei[a]，YU Long[b]，FENG Guanjun[c]，AISHAN Wumaier[d]，LI Pu[e]，ZHAO Jianguo[c]

（a. School of Software；b. Net Center；c. College of Humanities；d. College of Information Science and Engineering；e. School of Languages；Xinjing University，Urumqi 830046，China）

【Abstract】 Based on the characteristics of Uyghur language，a method of identifying Uyghur language event coreference relationship based on Stacked Denoising Autoencoder（SDAE） is proposed. This paper divides the Uyghur events to the candidate event pairs，extracted the nine features，basic characteristics of the event，the trigger word and the event distance. At the same time，the word embedding is used to calculate the semantic similarity of Uyghur events trigger words，taking semantic similarity as one of the features. And then training SDAE model，using softmax to complete the identification task of Uyghur language event coreference relationship. Experimental results show that SDAE is more suitable for the identification task than Support Vector Machine（SVM），the shallow machine learning model，and the use of word embedding further enhances the identification performance.

【Key words】 coreference relationship；Uyghur language；semantic similarity；stacked denoising autoencoder；deep learning

DOI： 10.19678/j.issn.1000-3428.0047731

0 概述

事件关系识别是一种针对“任意事件之间是否存在逻辑关系”进行判定的浅层事件关系检测任务[1]。共指关系作为自然语言处理中一类重要的事件关系，其正确的识别有助于人们有效地结合文本上下文来理解事件的详细经过，补充和扩展语义信息。事件识别与抽取作为自动内容抽取（Automatic Content Extraction，ACE）等评测会议的评测任务之一，是事件关系识别的基础，事件识别与抽取的效果直接影响事件关系识别的效果，对事件的识别与抽取已有一些研究成果[2-4]，在事件关系领域的研究中，文献[5]提出一种事件影响因子的计算方法，通过建立事件关系图（Event Relationship Map，ERM）来进行事件关系的表示，依据事件要素及其关系实现事件推理；文献[6]针对同一话题下事件关系抽取

基金项目： 国家自然科学基金（61662074，61563051，61262064）；国家自然科学基金重点项目（61331011）；新疆自治区科技人才培养项目（QN2016YX0051）。

作者简介： 王淑媛（1995—），女，硕士研究生，主研方向为自然语言处理；田生伟（通信作者），教授、博士；禹 龙，教授；冯冠军、艾山·吾买尔、李 圃，副教授、博士；赵建国，副教授、硕士。

收稿日期： 2017-06-27 **修回日期：** 2017-07-27 **E-mail：** tianshengwei@163.com

任务提出一种基于核心词和实体推理的事件关系抽取方法,实现事件间关联性的识别。文献[7]基于条件随机场方法识别信号词,并融入跨事件理论,采用最大熵模型的分类算法实现新闻事件的时序关系识别。目前有关事件关系的研究还主要集中在对某种特定事件关系类型的判定与识别方面,并且对事件共指关系的识别研究相对较少。

随着事件关系研究的深入,关系抽取任务中充分利用特征信息是近些年研究者的工作重点之一,而深度学习具有良好的特征表示和学习能力。深度学习最早由文献[8]提出,已在自然语言处理领域得到了广泛应用。文献[9]利用降噪自动编码器来学习文本的压缩和分布式表示,实现了中文新闻文本分类,与支持向量机(Support Vector Machine,SVM)相比,取得了更优的分类效果;文献[10]提出一种基于稀疏自编码模型对英文电子病历的实体关系抽取的方法,通过深度学习来学习上下文的一个表示,从而发掘词之间的组合关系特征,对于实体关系抽取任务实现了一定的效果提升。

针对上述研究现状,本文结合维吾尔语事件基本属性、结构信息及利用 Word Embedding 富含语义信息特性,抽取出事件基本属性、结构信息、触发词间的语义相似度等特征,提出一种基于堆栈降噪自编码(Stacked Denoising AutoEncoder,SDAE)的维吾尔语事件共指关系识别方法。

1 维吾尔语事件相关定义

为方便理解本文所研究的维吾尔语事件共指关系识别方法,首先明确以下定义:

定义 1(事件) 指在特定的环境和时间下发生,由若干角色参与,表现出动作特征的一件事情[11]。

定义 2(事件触发词) 能清晰地表示文本中事件发生的词。维语文本中事件的触发词主要由动名词和动词短语构成,维语的动词短语形式结构复杂多样,含有丰富的语法特点,其特殊的附加成分(词尾词缀)可以有效帮助识别事件信息。如例 1 所示(维吾尔语的书写格式为从右向左,下文同)。

例 1:بىر ماشىنىسى سوقۇلدى.

(一面包车与货车发生追尾。)

在例 1 中,سوقۇلدى(追尾)为该事件的触发词,表示该事件为一个突发事件,其中,سوقۇلدى(追尾)中的词尾体现出该事件时态为过去时。

定义 3(泛指事件) 文本上包含事件触发词,通过联系其上下文信息,不能代表一个事件的发生。如例 2 所示。

例 2:...دۆلەتلىك يەر تەۋرەش

(据中国地震网测定…)

在例 2 中,يەر تەۋرەش(地震)作为一个事件触发词,通过上下文信息分析,该触发词不能表示该句为一个事件,故将其当做一个泛指事件。

定义 4(事件极性) 结合 ACE(Automatic Content Extraction)语料库中对事件极性的定义以及实验组维吾尔语语言专家的意见,将事件极性分为 Positive 与 Negative2 类。根据事件的上下文信息,若明确指出该事件已经发生或正在发生,则该事件的极性为 Positive,否则为 Negative。

定义 5(语义类别) 语言学中将语义定义为语言形式和言语形式所表现出来的全部意义,根据实验组维吾尔语语言学专家的意见,将维语中语义类别划分为"rel_关系"、"time_时间"等 14 个类别。

定义 6(共指关系) 在维吾尔语文本中,当2 个事件指向同一个事件本体,则认为这 2 个事件具有共指关系。当多个事件指向同一个事件本体,认为这多个事件构成一个共指事件链,共指事件链中的任意 2 个事件都具有共指关系。如例 3 所示。

例 3:مەلۇم جايدا قاتناش ۋەقەسى يۈز بېرىپ،

بىر مىنىبوس بىلەن كىرا ماشىنىسى سوقۇلدى.

نۆۋەتتە ۋەقە سەۋەبى تەكشۈرىلىۋاتىدۇ.

(某地发生一起车祸(E1),一面包车与货车发生追尾(E2)。目前事故(E3)原因正在调查。)

在例 3 中,قاتناش ۋەقەسى(车祸)、سوقۇلدى(追尾)、ۋەقە(事故)3 个事件触发词都表示同一个突发事件:交通事故。则 E1、E2、E3 任意 2 个事件具有共指关系,E1、E2、E3 构成一个事件共指链。

2 基于 SDAE 模型的维吾尔语事件共指关系

本文研究的基本思想是:以篇章为单位,篇章中的维吾尔语事件为研究对象,将篇章中存在的任意事件两两组成候选事件对;结合维吾尔语语言特点,提取事件的基本属性、触发词、事件距离等结构特征,引入 Word Embedding,计算得到的触发词语义相似度作为特征之一,构成事件对的特征集合;利用该特征集合训练 SDAE 模型;通过 softmax 分类器对维吾尔语事件是否存在共指关系进行判定。

图 1 描述了整个处理流程。

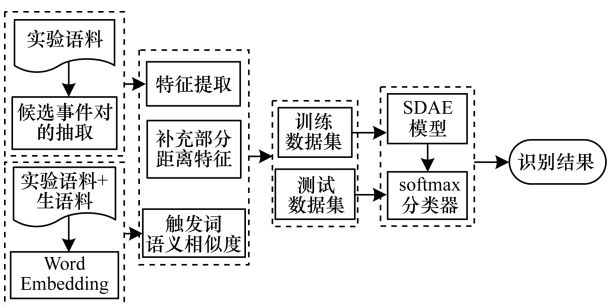

图 1 维吾尔语事件共指关系识别框架

2.1　候选事件对的抽取

候选事件对的抽取是进行后续实验的基础,本文首先对实验语料进行候选事件对的抽取,按照一定规则两两组成候选事件对,具体算法步骤如下:

步骤 1　提取出文本中的所有事件 e,存入集合 $\{eventSet\}$ 中,即 $e \in \{eventSet\}$。

步骤 2　遍历 $\{eventSet\}$,判断集合中每个元素 e 是否为泛指事件,若是,则将 e 从 $\{eventSet\}$ 中删去。

步骤 3　循环遍历集合 $\{eventSet\}$,依次取出事件元素 e,判断 e 对应的共指事件是否为空,不为空将 e 的共指事件 $e1$ 放入 $\{map\}$ 中,作为 *key* 值;将 e 以及和 $e1$ 具有共指关系的事件放入 $\{eventChain\}$ 集合中,作为 $\{map\}$ 中的 *value* 值;其中,$\{map\}$ 中的 *key* 和 *value* 值构成一个事件共指链。

步骤 4　将步骤 3 中得到的 *key* 值与 *value* 值中每个元素、以及 *value* 中的每个元素两两组合成事件对,作为实验的正例。

步骤 5　如果 $\{eventSet\}$ 中事件元素 e 对应的共指事件为空,那么将 e 与 $\{eventSet\}$ 中剩余元素两两组成事件对,作为实验的负例。

步骤 6　重复步骤 3 ~ 步骤 5,直到 $\{eventSet\}$ 为空;将步骤 4 与步骤 5 中得到的所有事件对作为实验的候选事件对。

2.2　Word Embedding 介绍

本文使用文献[12]提出的 Word2vec 工具训练得到 Word Embedding,选择 CBOW 模型作为训练框架。实验采用网络中无标注的维吾尔语语句,以及原有实验语料共同组成训练语料,通过对其进行去噪、去重处理,进而更好地获取每个触发词在低维空间中的语义分布情况。如例 4 所示。

例 4:مەلۇم جايدا قوراللىق ئېتىش ۋەقەسى يۈز بەردى،

ساقچى تەرەپنىڭ خەۋىرىگە قارىغاندا، بىر ئەر ئېتىپ ئۆلتۈرۈلدى.

(某地发生一起枪击(E1)事件。据警方报道,一名男子被击中(E2)。)

在例 4 中,事件 E1 和 E2 的触发词分别为 قوراللىق ئېتىش (枪击)、ئېتىپ ئۆلتۈرۈلدى (击中),对应的 Word Embedding(以 10 维为例)分别为:

قوراللىق ئېتىش (枪击) = [-0.087 2 0.078 7 -0.083 5 0.419 6 -0.467 2 0.522 5 0.109 1 0.321 0 0.763 1 0.008 6]

ئېتىپ ئۆلتۈرۈلدى (击中) = [0.015 4 -0.103 4 -0.300 1 0.573 8 -0.369 3 0.310 0 0.178 1 0.505 2 0.671 4 0.508 1]

通过计算触发词对应 Word Embedding 的余弦值来表示触发词间的语义相似度。利用式(1)计算得到 E1 与 E2 触发词的语义相似度为 0.850 4,将该值作为实验选取的特征之一。

$$\cos(\theta) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\| \boldsymbol{a} \| \times \| \boldsymbol{b} \|} = \frac{\sum_i (x_i \times y_i)}{\sqrt{\sum_i (x_i)^2} \times \sqrt{\sum_i (y_i)^2}} \tag{1}$$

其中,θ 为向量 $\boldsymbol{a}$、$\boldsymbol{b}$ 的夹角,$\boldsymbol{a} = [x_1, x_2, \cdots, x_n]$,$\boldsymbol{b} = [y_1, y_2, \cdots, y_n]$。

2.3　特征提取

提取有效的特征能提高 SDAE 模型对维吾尔语事件共指关系的识别性能。结合维吾尔语事件的特点,本文利用事件基本属性、结构特征等构成特征集合,作为 SDAE 模型的输入进行预训练。通过以下实例对实验所提取的特征进行介绍,具体如例 5 所示。

例 5:هايتىنىڭ شىمالى قىسمىدا 8 - چېسلا ئېغىر قاتناش ۋەقەسى يۈز بەردى، هايتى دۆلەتلىك جىددى

قۇتقۇزۇش مەركىزىنىڭ دوكلاتىغا قارىغاندا،

بۇ قېتىملىق قاتناش ۋەقەسى سەككىز ئادەمنىڭ ئۆلۈشى

ۋە 15 ئادەمنىڭ يارلىنىشىنى كەلتۈرۈپ چىقاردى،

ۋەقە سەۋەبىي ئىككى ماشىنىنىڭ ئۆز-ئارا سوقۇلۇشىدىن بولدى.

(海地北部山区 8 日发生一起严重车祸(E1)。据海地国家急救中心报道(E2),该事故(E3)造成 8 人死亡(E4),15 人受伤(E5)。是因为两辆车相撞(E6)造成的。)

分析如下:

1)事件的时态:同 ACE 语料库对时态的划分相同,将事件时态分为 Past(过去)、Present(现在)、Future(将来)、Unspecified(不确定)4 类,若候选事件对的时态一致,则特征值设置为 1,否则为 0;例 5 中,事件 E1 与 E6 时态均为 Past,故特征值为 1。

2)事件极性:若事件对的事件极性相同,设置特征值为 1,否则为 0;根据第 1 节中定义 4 得到例 5 中,事件 E1 与 E6 时态均为 Positive,故特征值为 1。

3)触发词的词性:在维吾尔语语言中,有 nn(普通名词)、nl(地点名词)、ad(形动词)、fd(副动词)等词性,根据事件触发词词性,判定词性相同设置为 1,否则为 0;例 5 中,事件 E1 词性为 nn_普通名词,E6 为 fd_助动词,E1 与 E6 触发词词性不同,故特征值为 0。

4)触发词的语义类别:根据触发词的语义类别是否相同,相同设置为 1,否则设置为 0;根据第 1 节中定义 5 得到例 5 中,事件 E1 与 E6 语义类别均为 event_事件,其语义类别相同,故特征值为 1。

5)触发词的类型:根据触发词的类型,将事件标注为本句事件、非本句事件和泛指事件。触发词类型一致,其特征值为 1,否则为 0;例 5 中,事件 E1 为本句事件,E6 为非本句事件,故特征值为 0。

6)事件对中触发词的相对距离:判断两个事件触发词的间隔词语数是否在给定区间内,是为 1,否则为 0。经统计语料库中的事件,相对距离在[0,10]所占比例约为 8%,[10,50]事件对所占比例约为 83%,其余占 9%,故本次实验将触发词的相对距离范围设置在[10,50];例 5 中,根据计算,事件 E1 与 E6 触发词的相对距离在该范围内,故特征值为 1。

7)事件对间隔句子数:从句法角度分析,一般具有共指关系的 2 个事件具有跨句子、跨段落的特点。

故本文将事件对间隔句子差在[0,5]内的特征值设置为1,否则为0;例5中,E1与E6间隔1句,该事件对特征值为1。

8)事件对的间隔事件数:根据统计,2个事件间隔的事件数在[2,5]范围内共指所占比例为80%,将事件对在[2,5]范围内的特征值取1,否则取0;例5中,E1与E6间隔4个事件,故特征值为1。

9)事件对是否具有依存关系:根据句法分析得到的依存关系表来确定,关联事件对必须在同一个句子中,否则不具有依存关系。将在同句中的特征值设置为1,否则为0。例5中,E1与E6不在同一个句子中,故特征值为0。

10)事件对中触发词的语义相似度:根据2.2节介绍,利用触发词的Word Embedding所表示的语义信息,计算触发词的语义相似度,根据计算,例5中E1与E6对应的触发词的语义相似度为0.678 5。

2.4 堆栈降噪自编码

堆栈降噪自编码(Stacked Denoising Autoencoder,SDAE)是深度学习中一种重要的模型,已广泛应用于图像分类、行为识别、故障诊断等领域[13-15]。

SDAE模型具有良好的特征学习能力,该模型能从原始数据集中学习到更抽象且高度区分的特征表示,最终将从模型中学习到的特征作为softmax分类器的输入从而实现分类。SDAE模型由多个降噪自编码(Denoising Autoencoder,DAE)逐层堆叠构成,该模型中前一个编码器的隐层输出作为其后一层编码器可视层的输入,如图2所示。

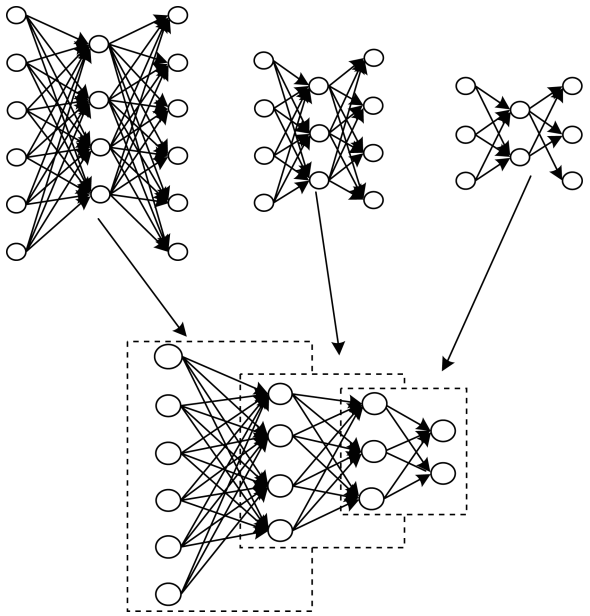

图2 堆栈降噪自编码模型

SDAE训练过程分为预训练和微调2个阶段,首先采用逐层贪婪训练法完成无监督预训练,在预训练结束后,采用反向传播算法,利用梯度下降方法对所有层的参数进行微调,进而优化整个模型。

2.5 降噪自编码

自编码器(AutoEncoder,AE)采用一种无监督的学习算法,包括编码和解码两部分。它尝试学习一个恒等函数,使得输出值接近于输入值。

AE在学习过程中仅仅简单地保留原始输入样本的信息,从而不能确保对输入样本提取出一种有用的特征表示。为避免这种情况,DAE在AE的基础上,通过将训练数据加入噪声,使得AE必须学习去除这种噪声从而获得没有被噪声污染过的输入数据,如图3所示。因此,DAE学习的特征更具鲁棒性,提高了模型的泛化能力。

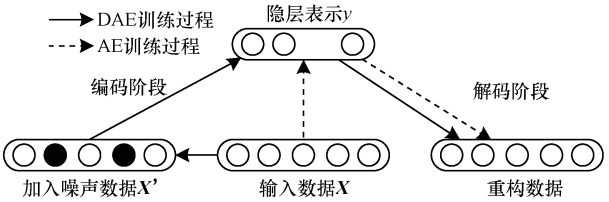

图3 AE与DAE训练过程

将2.3节中从维吾尔语事件中得到的特征向量 $\boldsymbol{X}\in[0,1]^N$ 作为数据输入层,DAE通过一个随机的映射变换 $\boldsymbol{X}\sim q_N(\boldsymbol{X}'|\boldsymbol{X})$,使得原始输入数据 $\boldsymbol{X}$ 被"破坏",得到含有噪声的数据向量 $\boldsymbol{X}'\in[0,1]^N$,则DAE中编码器的输出如下:

$$\boldsymbol{Y}'=f_\theta(\boldsymbol{X})=s(\boldsymbol{W}\boldsymbol{X}'+\boldsymbol{b}) \tag{2}$$

通过解码过程将激活值向量 $\boldsymbol{Y}'\in[0,1]^N$ 反向变换为对原始输入 $\boldsymbol{X}$ 的重构表示向量 $\boldsymbol{Z}\in[0,1]^N$,如下:

$$\boldsymbol{Z}=g_{\theta'}(\boldsymbol{Y}')=s(\boldsymbol{W}'\boldsymbol{Y}'+\boldsymbol{b}) \tag{3}$$

其中,$\theta=\{\boldsymbol{W},\boldsymbol{b}\}$ 为编码参数集合,$\boldsymbol{W}\in R^{(M\times N)}$ 为权重矩阵,$\boldsymbol{b}\in R^M$ 为偏置矩阵,$\theta'=\{\boldsymbol{W}',\boldsymbol{b}\}$ 为解码参数,激活函数 $s(x)=\dfrac{1}{(1+\exp(-x))}$,值域为[0,1]。

最后通过不断调整 θ 与 θ' 的值,得到最小化重构误差 J,重构误差可表示为:

$$J=\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\|z_i-x_i\|^2\right)\right]+\frac{\lambda}{2}\sum_{l}\sum_{i}\sum_{j}(w_{ij})^2 \tag{4}$$

其中,λ 为权重衰减项系数,第一项表示均方误差,目的是为了最小化重构误差,第二项表示权重衰减项,目的是为了减小权重的幅度防止出现过拟合现象。

3 实验结果与分析

实验采用准确率 P、召回率 R 以及 F 值作为实验性能的评价指标。其中,准确率 P 是指正确识别的事件对共指关系数占实际识别的百分比,R 是指正确识别的事件对共指关系数占系统中存在的事件对共指关系数的百分比。F 是准确率和召回率的综合评价指标,即 $F=P\times R\times 2/(P+R)$。为避免实验

的偶然性和随机性,实验均采用 10 折交叉验证,取平均值作为最终的实验结果。

3.1　基于 SDAE 模型的有效性验证

3.1.1　SDAE 模型中不同隐层层数对实验的影响

在不引入触发词语义相似度特征的前提下,使用 2.3 节提取的前 9 个特征作为模型输入,验证具有不同隐层层数的 SDAE 模型对维吾尔语事件共指关系的影响。SDAE_i 中 i 表示 SDAE 包含的隐层层数,实验结果如表 1 所示。

表 1　基于 SDAE 模型的有效性验证结果　%

模型	*P*	*R*	*F*
SDAE_1	79.90	70.97	73.69
SDAE_2	80.18	71.21	74.02
SDAE_3	80.16	74.64	76.58
SDAE_4	82.13	69.29	73.77

由表 1 可知,不同隐层层数的 SDAE 模型在维吾尔语事件共指关系识别任务中识别性能不同,隐层层数为 3 时,*F* 为 76.58%,识别性能达到最优。而隐层层数较低时,模型挖掘数据深层语义的能力较弱,未能充分学习数据中隐含的语义特征;当隐层层数为 4 时,*F* 的值开始下降,其原因是出现过拟合现象,模型的泛化能力降低。实验结果表明,SDAE 模型能有效的运用在维吾尔语事件共指关系的识别任务中,并且在本文实验中隐层层数为 3 时取得较优效果。

3.1.2　SDAE 模型与 SVM 的对比实验

根据 3.1.1 节结果,本次实验选择性能最优的 SDAE_3 与 SVM 进行对比实验。SVM 是一种典型的处理非线性数据的浅层机器学习模型,与采用深度学习机制的 SDAE 模型有较好的可比性。实验结果如表 2 所示。

表 2　SDAE_3 与 SVM 对比实验结果　%

模型	*P*	*R*	*F*
SVM	78.05	65.56	68.77
SDAE_3	80.16	74.64	76.58

由表 2 可知,SDAE_3 中 3 个评价指标的结果均高于 SVM,*P* 提升了 2.11%,*R* 提升了 10.08%,*F* 提升了 7.81%。这是因为深度学习机制对有限的特征具有良好的自学习能力,能学习到更抽象更具有识别意义的特征。实验结果表明,相较于 SVM 模型,采用深度学习机制的 SDAE 模型在维吾尔语事件共指关系的识别任务中性能更好。

3.2　Word Embedding 对实验的影响

3.2.1　Word Embedding 对 SDAE 模型的影响

Word Embeding 富含丰富的语义信息,为探讨 Word Embedding 对维吾尔语共指关系识别性能的影响,选用 100 维的 Word Embedding 计算触发词语义相似度进行实验,依次训练不同隐层层数的 SDAE 模型。实验结果如表 3 所示。

表 3　Word Embedding 的引入对实验的影响　%

模型	*P*	*R*	*F*
SDAE_1	79.90	70.97	73.69
SDAE_1 + W_E	81.34	76.02	77.97
SDAE_2	80.18	71.21	74.02
SDAE_2 + W_E	82.46	74.05	77.39
SDAE_3	80.16	74.64	76.58
SDAE_3 + W_E	82.48	75.28	78.10
SDAE_4	82.13	69.29	73.77
SDAE_4 + W_E	84.80	69.65	76.66

由表 3 可知,对于不同隐层层数的 SDAE 模型,Word Embedding 的引入对 3 个评价指标在一定程度上都有所提升。例如,SDAE_1 + W_E 同 SDAE_1 相比,其 *P* 增加了 1.44%,*R* 增加了 5.05%,*F* 增加了 4.28%。这是因为 Word Embedding 对自然语言有良好的表征能力,促进了 SDAE 模型对特征更深层次的学习。实验结果表明,通过引入 Word Embedding 计算得到的语义相似度特征,能提升模型对维吾尔语事件共指关系的识别性能。

3.2.2　Word Embedding 维度对 SDAE 模型的影响

Word Embedding 的不同维度计算得到的触发词语义相似度不同,对 SDAE 模型的性能有一定的影响。本次实验探索 Word Embedding 维度设置对表达触发词语义信息的影响,将维度分别设定为10 维、50 维、100 维、150 维、200 维。根据表 1 选定性能最优的 SDAE_3 进行本次实验,结果如表 4 所示。

表 4　Word Embedding 维度对实验的影响　%

维度	*P*	*R*	*F*
10	82.82	65.38	71.18
50	82.79	66.78	72.24
100	82.48	75.28	78.10
150	84.84	72.77	76.62
200	81.74	68.15	72.71

由表 4 可知,在 SDAE_1 中加入由不同维度的 Word Embedding 计算得到的触发词语义相似度,均提高了识别效果的准确率,根据 3 个评价指标综合比较,将 Word Embedding 的维度设置为 100 维时,其计算得到的触发词语义相似度的值,使实验获得最优效果,*F* 达到 78.10%。原因是当维度过高时,Word Embedding 虽然包含丰富的语义信息,但也引入了噪音及不必要的干扰信息,影响触发词语义相似度的计算结果,从而影响维吾尔语事件共指关系的识别性能。

通过本文 2 个实验可以看出,在特征中引入由 Word Embedding 计算得到的触发词语义相似度,能有效地提高维吾尔语事件共指关系识别效果。

4 结束语

维吾尔语事件共指关系识别对维吾尔语自然语言领域研究的发展有重要的意义,将其结合深度学习机制能更好地挖掘文本中深层语义信息的优势。为此,本文探索了深度学习方法在维吾尔语事件共指识别的应用,并通过实验证明了其有效性。后期将利用训练词向量进行事件表示来减少特征工程的使用,避免人工干预,进一步提升识别性能。

参考文献

[1] 马 彬.事件关系识别关键技术研究[D].苏州:苏州大学,2014.

[2] 赵妍妍,秦 兵,车万中,等.中文事件抽取技术研究[J].中文信息学报,2008,22(1):3-8.

[3] LIU Mengyi, LIU Xin, LI Yan, et al. Exploiting feature hierarchies with convolutional neural networks for cultural event recognition[C]//Proceedings of IEEE International Conference on Computer Vision Workshop. Washington D. C., USA: IEEE Computer Society, 2015: 274-279.

[4] NGUYEN T H, CHO K, GRISHMAN R. Joint event extraction via recurrent neural networks [C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Washington D. C., USA: 2016: 300-309.

[5] 仲兆满,刘宗田,周 文,等.事件关系表示模型[J].中文信息学报,2009,23(6):56-60.

[6] 杨雪蓉,洪 宇,马 彬,等.基于核心词和实体推理的事件关系识别方法[J].中文信息学报,2014,28(2):100-108.

[7] 丁 础,周 枫,庙介璞,等.基于跨事件理论的新闻事件时序关系识别方法[J].计算机工程,2017,43(6):189-194.

[8] 余 凯,贾 磊,陈雨强,等.深度学习的昨天、今天和明天[J].计算机研究与发展,2013,50(9):1799-1804.

[9] 刘红光,马双刚,刘桂锋.基于降噪自动编码器的中文新闻文本分类方法研究[J].现代图书情报技术,2016,32(6):12-19.

[10] 吴嘉伟,关 毅,吕新波.基于深度学习的电子病历中实体关系抽取[J].智能计算机与应用,2014,4(3):35-38.

[11] 付剑锋.面向事件的知识处理研究[D].上海:上海大学,2010.

[12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[13] LIAN Peng, SHI Wenzhong, ZHANG Xiaokang. Remote sensing image classification based on stacked denoising autoencoder[J]. Remote Sensing, 2017, 10(1): 16.

[14] BUDIMAN A, FANANY M I, BASARUDDIN C. Stacked denoising autoencoder for feature representation learning in pose-based action recognition[C]//Proceedings of the 3rd IEEE Global Conference on Consumer Electronics. Washington D. C., USA: IEEE Press, 2014: 684-688.

[15] LU Chen, WANG Zhenya, QIN Weili, et al. Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification [J]. Signal Processing, 2017, 130(C): 377-388.

编辑 刘 冰

(上接第 304 页)

[6] YE M, YIN P, LEE W C. Location recommendation for location-based social networks[C]//Proceedings of the 18th International Conference on Advances in Geographic Information Systems. New York, USA: ACM Press, 2010: 458-461.

[7] 巫 可,战荫伟,李 鹰.融合用户属性的隐语义模型推荐算法[J].计算机工程,2016,42(12):171-175.

[8] 何 顺,王淑娟,雷建云.基于加权多融合偏好与结构相似度的协同过滤算法[J].计算机工程,2016,42(10):64-68.

[9] LIU B, FU Y J, YAO Z J, et al. Learning geographical preferences for point-of-interest recommendation[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2013: 1043-1051.

[10] YANG D, ZHANG D, YU Z, et al. A sentiment-enhanced personalized locaiton recommendation system [C]//Proceedings of the 24th ACM Conference on Hypertext and Social Media. New York, USA: ACM Press, 2013: 119-128.

[11] BAO J, ZHENG Y, MOKBEL M F. Location-based and preference-aware recommendation using spares geo-social networking data [C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems. New York, USA: ACM Press, 2012: 199-208.

[12] JANARTHANAN M, GANDHI M. Location based preference aware recommendation using sparse geo-social networking data[J]. European Journal of Applied Sciences, 2016, 8(3): 181-185.

[13] LI Q, ZHENG Y, XIE X, et al. Mining user similarity based on location history [C]//Proceedings of International Conference on Advances in Geographic Information Systems. New York, USA: ACM Press, 2008: 34-38.

[14] GAO H, TANG J, HU X, et al. Exploring temporal effects for location recommendation on location-based social networks [C]//Proceeding of the 7th ACM Conference on Recommender Systems. New York, USA: ACM Press, 2013: 93-100.

[15] YUAN Q, CONG G, MA Z, et al. Time-aware point-of-interest recommendation [C]//Proceeding of ACM SIGIR International Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2013: 363-372.

[16] ZHANG J D, CHOW C Y. TICRec: a probabilistic framework to utilize temporal influence correlations for time-aware location recommendations[J]. IEEE Transactions on Services Computing, 2016, 9(4): 633-646.

编辑 吴云芳