# 注意力机制的 LSTM-DBN 维语人称代词指代消解

李东欣[1],禹　龙[2],田生伟[1],李　圃[3],赵建国[4]
(1. 新疆大学 软件学院,新疆 乌鲁木齐 830008;
2. 新疆大学 网络中心,新疆 乌鲁木齐 830008;
3. 新疆大学 语言学院,新疆 乌鲁木齐 830046;
4. 新疆大学 人文学院,新疆 乌鲁木齐 830046)

**摘　要**:针对维吾尔语中人称代词指代歧义问题,结合维吾尔语言的词法、语法、词间位置等关系,以及注意力机制、长短时记忆网络和深度置信网络,提出了一种维语人称代词指代消解模型。首先,分析维语中人称代词指代的特点和表达规律,提取出相应词向量特征;其次,借助长短时记忆网络挖掘维吾尔语人称代词的语义特征,并利用注意力机制的相似性度量、权重调节能力,避免信息在层间传递的丢失,实现特征编码向量的信息整合;最后利用深度置信网络(DBN)进一步挖掘出隐藏在维语上下文中的深层语义特征,完成维语人称代词指代消解。实验结果表明,所提模型在挖掘深层语义信息和识别效果上优于传统的深度学习模型,准确率达到了 81.14%,$F_1$ 达到了 78.83%。

**关键词**:人称代词;指代消解;词向量;注意力机制;深度信念网络;长短时记忆网络
**中图分类号**:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2019)07-0033-06
doi:10.3969/j.issn.1673-629X.2019.07.007

# Attention Mechanism of LSTM-DBN Uyghur Personal Pronoun Anaphora Resolution

LI Dong-xin[1], YU Long[2], TIAN Sheng-wei[1], LI Pu[3], ZHAO Jian-guo[4]
(1. School of Software, Xinjiang University, Urumqi 830008, China;
2. Network Center, Xinjiang University, Urumqi 830008, China;
3. School of Languages, Xinjiang University, Urumqi 830046, China;
4. School of Humanities, Xinjiang University, Urumqi 830046, China)

**Abstract**: Aiming at the anaphora ambiguity of personal pronouns in Uyghur language, combining with the lexical, grammar and inter-word position of Uyghur language as well as the attention mechanism, long-short term memory and deep belief networks, an anaphora resolution model is proposed. Firstly, the characteristics and expression patterns of the personal pronouns in Uyghur language is analyzed and the corresponding word vector characteristics is extracted. Secondly, long short-term memory network is used to explore the semantic features of personal pronoun and the attention mechanism of similarity measurement and weights adjustment ability is used to avoid the loss of the information transfer between layers, to realize the information integration of the feature encoding vector. Finally, the deep belief network (DBN) is applied to further explore the deep semantic features hidden in Uyghur context and complete the Uyghur personal pro-nouns anaphora resolution. The experiment shows that the proposed model is superior to the traditional deep learning model in the mining of deep semantic information and recognition effect, with an accuracy of 81.14% and 78.83% of $F_1$ respectively.

**Key words**: personal pronouns; anaphora resolution; word embedding; attention-based mechanism; deep belief- network; long short-term memory

**收稿日期**:2018-09-01　　**修回日期**:2019-01-09　　**网络出版时间**:2019-03-21
**基金项目**:国家自然科学基金(61563051,61662074,61262064);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);新疆天山青年计划项目(2017Q011)
**作者简介**:李东欣(1991-),男,研究生,研究方向为自然语言处理;禹　龙,硕士,教授,通讯作者,研究方向为自然语言处理;田生伟,博士,教授,研究方向为自然语言处理;李　圃,博士,副教授,研究方向为维汉双语和应用语言学;赵建国,博士,副教授,研究方向为维汉双语对比。
**网络出版地址**:http://kns.cnki.net/kcms/detail/61.1450.TP.20190321.0942.076.html

## 0 引 言

在篇章级别文本语义的整体理解上,准确没有歧义的指代消解对其具有很大的影响。在信息抽取、自动文摘等自然语言处理中具有重要的作用[1]。Mc-Carthy等[2]将其转换为二分类问题,用于判断先行语和照应语之间的指代关系。王荣波等[3]基于篇章级别设计的多元判别分析模型,提高了句群自动划分的精确度。李国臣等[4]利用机器学习算法结合优先选择策略,针对篇章级别的文本,进行了指代消解研究。Ng 等[5]研究了在挖掘语义信息方面指代消解所起的作用。Kong 等[6]探索了更深层次的语义信息对指代消解的影响。许敏等[7]采用了格框架的方法进行指代消解。之后,王厚峰等[8-9]在中文领域给出了消解人称代词的基本规则。董国志等[10]提出了将语料库、规则预处理和最大熵模型相结合的方法。王海东[11]和孔芳[12]将语义角色应用在指代消解模型中,实验结果显示,引入语义角色能够更好地提高消解模型的准确率。

上述研究尽管在一定程度上提高了指代消解模型的性能,但是需要人工参与进行特征抽取和分析,因此,仍然存在许多不足。如:过程繁琐,耗时太久;传统浅层学习方法不能够很好地挖掘文本中深层的语义信息;处理复杂问题时,常常会出现泛化能力不足现象;不能很好地挖掘语义的深层细节信息。针对上述问题,文中利用注意力机制、长短时记忆网络和深度信念网络,构建了一种维吾尔语的人称代词指代消解模型。

## 1 相关研究

随着 attention 机制和深度学习算法在图像处理、目标检测和语音、视频识别等众多领域的广泛应用,也为指代消解的研究提供了全新的思路[13]。

Collobert[14]将词汇向量化,并作为初始值来训练指代消解模型。胡乃全[15]将特征向量应用在中文人称代词指代消解中,有效提高了系统的性能。Hinton[16]提出了基于 RBM 的 Log-Bilinear 语言模型。Hochreiter 等利用长短记忆单元(long short-term memory)[17]模型有效解决了传统的 RNN 训练时的梯度爆炸和梯度消失问题,让 RNN 能真正有效地利用长短距离的信息;胡新辰等[18]将 LSTM 模型应用于语义关系分类问题,并取得了很好的效果。随后 attention 机制被大量应用于各种图像处理和自然语言处理模型中,为进一步解决传统 attention 机制的局限性,文献[19]将 attention 机制和 RNN 模型相结合,并提出全局(global)机制和局部(local)机制。文献[20]利用 attention-based 得到含有输入序列节点注意力概率分布的语义编码,并将其作为分类器的输入,以缓解特征向量提取过程中的信息丢失和信息冗余等问题。

## 2 维吾尔语人称代词的特点

维吾尔语是一种黏着型语言,它的不同语法形式主要是通过在词语的词尾处添加不同的词缀来体现的。例如:قالدى ئويلىنىپ،يەنە تۇرسۇن(吐尔逊觉得),词缀“قالدى ئويلىنىپ”缀接到人名“ئويلىنىپ،يەنە تۇرسۇن(吐尔逊)”后,表达“吐尔逊觉得”的意思。

人称代词在维语中的形式与汉语、英语有着明显的区别。(1)前者人称代词不包括反身代词,而英语和汉语包括反身代词;(2)维语的第三人称代词不仅没有性别区分,还可以指物体;(3)一、二人称有单复数之分,而第三人称没有。因此维语人称代词的单复数特征,为指代消解的研究提供了一个很好的依据。

## 3 基于 Attention-Based LSTM-DBN 的人称代词指代消解

### 3.1 人称代词指代消解整体处理流程

文中结合 attention 机制、LSTM 模型和深度信念网络实现维吾尔语人称代词指代消解。其基本思想是:首先确定先行语和照应语对应的候选项,构建人称代词特征向量,挖掘出人称代词语义信息;然后利用多层感知器将十一项规则特征与挖掘出的人称代词语义信息进行融合;最后由 softmax 分类器进行分类,完成消解任务。指代消解整体流程如图 1 所示。

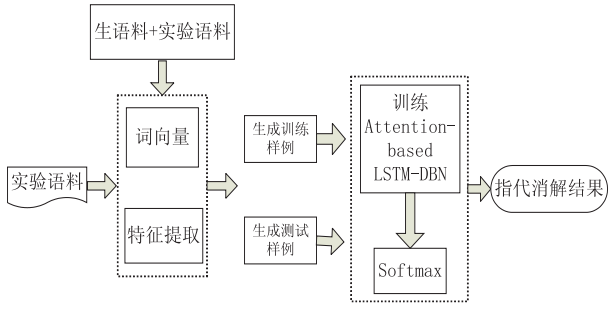


图 1 基于 Attention-Based LSTM-DBN 的维吾尔语人称代词指代消解框架

### 3.2 维语人称代词指代消解模型

通过 Attention-Based LSTM 挖掘文本中照应语和候选先行语上下文的语义特征,并作为深度信念网络的输入;然后经过 DBN 进一步挖掘出隐藏在文本中的深层语义特征;最后将挖掘出的人称代词语义特征与特征规则融合,经过 softmax 进行分类,完成维吾尔语人称代词指代消解。

其中,$\boldsymbol{X}=\{X_1,X_2,\cdots,X_i\}$ 是文本词语序列的词向量;$H_i$ 是文本中人称代词经过 LSTM 模型后的输出,表示为 $H_i=\{h_1,h_2,\cdots,h_k\}$;$\boldsymbol{W}_i$ 是人称代词构成的词向量矩阵,表示为 $\boldsymbol{W}_i=\{W_1,W_2,\cdots,W_k\}$;$\theta$ 表示注意力概率权重;$r_i$ 是 DBN 模型的输入数据。$V_i=\{V_0,V_1\}$ 表

示受限玻尔兹曼机中的显性神经元; $H_i^* = \{h_0^*, h_1^*\}$ 表示受限玻尔兹曼机中的隐性神经元; $W_k^* = \{W_0^*, W_1^*, W_2^*\}$ 表示它们之间连接的权重。图 2 是指代消解模型的具体框架。

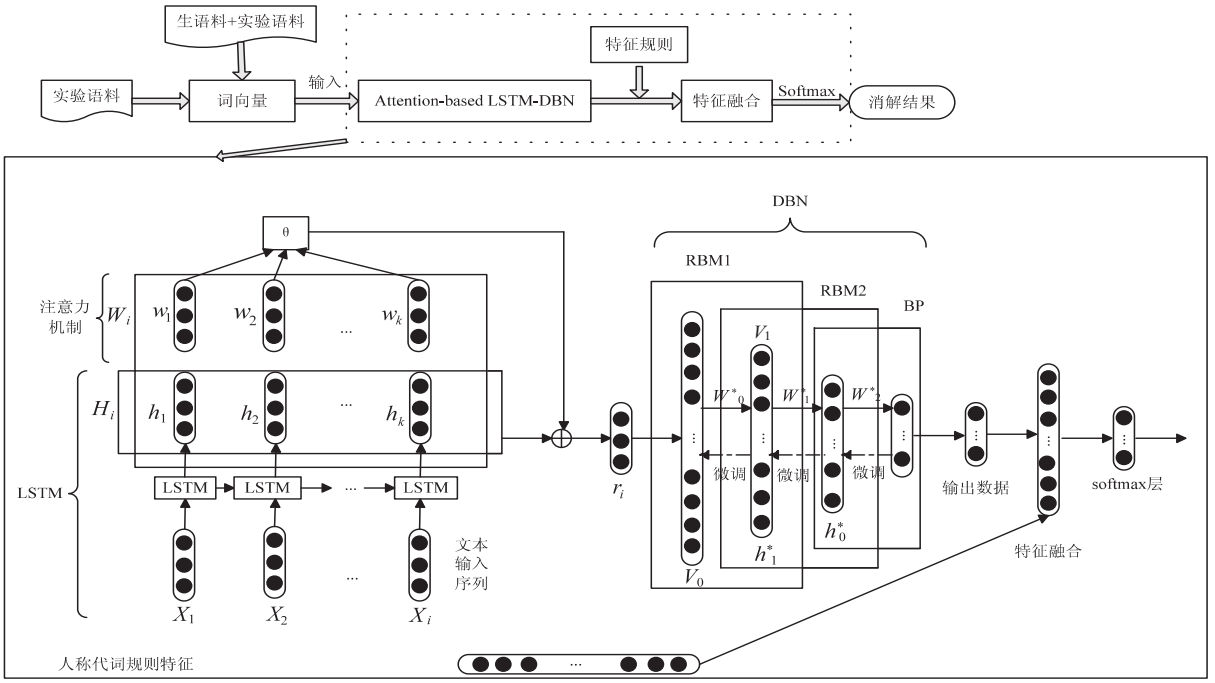


图 2　维吾尔语人称代词指代消解模型具体框架

### 3.3　注意力机制 Attention-Based 模型

针对在词汇转换成中间向量时,会导致很多细节信息缺失问题。文中通过添加注意力机制来提高模型输出信息的质量,减少计算时耗。

在图 2 中 $\theta$ 就是历史节点对最后节点的注意力概率, $\boldsymbol{X}_i$ 是文本词语向量表示。计算出 $\boldsymbol{X}_i$ 对于文章总体的影响力权重,可突出关键词的作用,减少非关键词对于文本整体语义的影响。文中在编码阶段使用 Attention-Based 机制。维吾尔语人称代词语义特征表达式为:

$$\theta = \sum_{i=1}^{K} W_{ik} h_i \tag{1}$$

$$r_i = \tanh(W_x h_m + W_p r_j) \tag{2}$$

$$r_j = \theta H_i \tag{3}$$

语义编码 $\theta$ 主要是通过注意力概率权重与历史输入节点的隐藏层的状态乘积的累加得到,表示人称代词经过模型后的语义表示; $K$ 表示输入序列的元素数目; $W_{ik}$ 表示节点 $K$ 对于节点 $i$ 的注意力概率权重; $\boldsymbol{W}_x$ 和 $\boldsymbol{W}_p$ 分别是模型训练时 $h_m$ 和 $r_j$ 的权重向量。

### 3.4　LSTM 模型

LSTM 模型是通过在 RNN 的基础上添加细胞控制机制(cell state),并通过输入门、遗忘门、输出门的控制,解决了 RNN 模型长期依赖问题和序列过长导致的梯度爆炸问题。

针对维吾尔语人称代词特征选择问题,文中采用结合注意力机制的 LSTM 模型用于提取特征。传统的模型在挖掘文本语义信息时,往往忽略了上下文语义信息,使得信息缺失严重。LSTM 模型具有短暂的记忆存储功能,在挖掘人称代词语义信息时可以充分利用记忆单元中存储的上一时刻的词汇信息,挖掘出当前时刻人称代词的语义特征;因此 LSTM 模型能够更好地从上下文中挖掘出人称代词的语义信息。

设输入的词序序列为 $X = \{X_1, X_2, \cdots, X_i\}$ ,在 $t$ 时刻,LSTM 的输入有三个:(1)当前时刻 LSTM 的输入值 $x_t$ ;(2)上一时刻 LSTM 的输出值 $h_{k_{t-1}}$ ;(3)上一时刻的单元状态 $C_{k_{t-1}}$ 。LSTM 的输出也有两个:当前时刻 LSTM 的输出值 $h_{k_t}$ ;当前时刻的单元状态 $C_{k_t}$ 。则在 $t$ 时刻 LSTM 单元可以表述为:

$$f_{k_t} = \delta(W_f \cdot [h_{k_{t-1}}, x_t] + b_f) \tag{4}$$

$$i_{k_t} = \delta(W_i \cdot [h_{k_{t-1}}, x_t] + b_i) \tag{5}$$

$$O_{k_t} = \delta(W_o \cdot [h_{k_{t-1}}, x_t] + b_o) \tag{6}$$

$$h_{k_t} = O_{k_t} \cdot \tanh C_{k_t} \tag{7}$$

$$C_{k_t} = f_{k_t} C_{k_{t-1}} + i_{k_t} \delta(W_C [h_{k_{t-1}}, x_t] + b_C) \tag{8}$$

其中, $f$ , $i$ , $O$ , $C$ 分别表示模型中的遗忘门、输入门、输出门和记忆单元; $W$ 为权重; $b$ 为 LSTM 模型中的偏置项; $\delta$ 为激活函数 sigmoid。

### 3.5　深度置信网络

为了确保在训练过程中,特征向量映射到不同空间特征时,都尽可能多地保留特征信息,减小对学习目标过拟合的风险,文中在模型后半部分采用深度置信网络。

深度置信网络是由若干层受限玻尔兹曼机 RBM

和一层有监督的反向传播网络 BP 组成的。图 2 中 $V_i$ 和 $H_j^*$ 分别表示显性神经单元和隐性神经单元，$W_k^*$ 表示它们之间的连接权重，用来微调整个实验模型。受限玻尔兹曼机 RBM 是一种能量模型，能量函数定义为：

$$E(v,h^*) = -\sum_i a_i v_i - \sum_j h_j^* b_j^* - \sum_{ij} W_k^* v_i h_j^* \quad (9)$$

其中，$a_i$ 和 $b_j^*$ 分别是显性神经单元偏置项和隐藏神经单元偏置项。

训练过程可分为：

预训练：单独地无监督地训练每一层 RBM 网络，确保网络获得高阶抽象特征。

微调：利用反向传播网络微调网络的权重。

### 3.6 softmax 分类器

利用多层感知器将 Attention－Based LSTM－DBN 模型学习到的维吾尔语人称代词语义特征与人称代词特征规则进行融合，然后将融合后的特征送到 softmax 分类器进行分类，并明确照应语和先行语的指代关系，完成维吾尔语人称代词指代消解研究。

### 3.7 生成训练实例和测试实例

将人称代词与其之前出现的名词短语按照一定的规则进行两两配对。生成训练实例时，因为指代链的信息是已经知道的，所以可以先对已识别出的人称代词进行判断，确定其是否在某个指代链中。若在，则将其视为照应语，并查找该照应语对应的先行语；如果不存在，则将该人称代词视为非待消解项，而且不用寻找该人称代词对应的先行语。经过统计实验语料，在文中实验中，将距离某个照应语 $X_n$ 在五句之内的所有名词短语视为匹配项，并将匹配项与该照应语一一进行匹配。若是存在某个名词短语 $NP_i$ $(0<i<n)$，使得 $NP_i$ 与照应语 $X_n$ 之间存在指代关系，则将这对名词短语对视为正例；若不存在，则将其视为负例。

在生成测试实例时，因为指代链的信息都是未知的，所以将识别出的所有人称代词都视为照应语，与其距离为五句之内的名词短语依次进行匹配，配对形式为<照应语，候选先行语>，然后通过模型判断它们之间是否存在指代关系。

### 3.8 特征提取

不同的特征对模型的消解性能具有重要的影响。因此，提取的特征要能够使模型快速、有效、准确地对词汇间的指代关系进行判断。经过查看阅读国内外大量的关于汉语和英语指代消解的研究文献，结合维语特点通过实验筛选出以下十一个特征。

(1) 如果照应语是代词(Anaphor Pronoun.)：此特征表示为 $V_{ap}=\{0,1\}$，如果照应语是代词，则 $V_{ap}=1$；如果不是，则 $V_{ap}=0$。

(2) 如果候选先行语是代词(Candidate Pronoun.)：此特征表示为 $V_{cp}=\{0,1\}$，如果候选先行语是代词，则 $V_{cp}=1$；否则 $V_{cp}=0$。

(3) 是否嵌套(Nest Pron.)：此特征表示为 $V_{nest}=\{0,1\}$，如果照应语与候选先行语都是互相嵌套，特征值 $V_{nest}=1$；否则 $V_{nest}=0$。

(4) 性别一致性(Gender Agreement.)：该特征表示为 $V_{ga}=\{0,0.5,1\}$，如果照应语和候选先行语的性别一致，特征值 $V_{ga}=1$；如果性别不一致，则特征值 $V_{ga}=0$；如果照应语和候选先行语有一个未知，特征值 $V_{ga}=0.5$。

(5) 语义类别的一致性(Semantic Agreement.)：该特征表示为 $V_{sa}=\{0,0.5,1\}$，如果候选先行语与照应语的语义类别一致，该特征值 $V_{sa}=1$；如果不一致，则 $V_{sa}=0$；如果照应语和候选先行语中有一个未知，该特征值 $V_{sa}=0.5$。

(6) 单复数的一致性(Number Agreement.)：该特征表示为 $V_{na}=\{0,0.5,1\}$，如果照应语和候选先行语的单复数一致，该特征值 $V_{na}=1$；如果不一致，$V_{na}=0$；如果照应语和候选先行语中有一个未知，则该特征值 $V_{na}=0.5$。

(7) 词性的一致性(POS Agreement.)：该特征表示为 $V_{pos}=\{0,1\}$，如果候选先行语与照应语词性一致，该特征值 $V_{pos}=1$；否则 $V_{pos}=0$。

(8) 命名实体特征(Name Entity.)：该特征表示为 $V_{name}=\{0.1,0.3,0.6,1\}$，若候选先行语的实体类型是人名，该特征值取 1；若候选先行语的实体类型是机构名，该特征值取 0.3；若是地名，该特征值取 0.6；若是其他，该特征值取 0.1。

(9) 语义角色特征(Semantic Role.)：该特征表示为 $V_{role}=\{0,1\}$，若候选先行语的语义角色是施事者，则该特征值 $V_{role}=1$；否则 $V_{role}=0$。

(10) “格”语法一致性(Case Gramma.)：该特征表示为 $V_{cg}=\{0,0.5,1\}$，如果候选先行语和照应语格语法一致，则该特征值 $V_{cg}=1$；若不一致，则 $V_{cg}=0$；若照应语和候选先行语中有一个格语法未知，则该特征值 $V_{cg}=0.5$。

(11) 距离特征(Distance.)：该特征表示照应语和候选先行语语句的空间距离。距离越大，存在的指代关系的可能性越小。特征表示为 $V_{distance}=g(d)$，对空间距离进行逆向取值，并归一化在 0 和 1 之间。

设空间距离为 $d$，若 $d\geqslant 10$，则 $V_{distance}=1$；若 $d<10$，则 $V_{distance}=0.1\times(10-d)$。

例 1：قاراپ ئالدىغا ھالدا ئاچچىقلانغان ئالىم چىڭ تەرەپتىن ئارقا بولسا يۈگۈل ئممما، قوغلىدى ئۇنى ئۆتۈۋالدى (阿里木愤怒的向前追去，古丽却从后边紧紧地拽住了他。)

人称代词“ئۇنىــ(他)”,与该人称代词前面的名词短语匹配,然后规则过滤,“گۈلىــ(古丽)”和“ئۇنىــ(他)”构成了反例<ئۇنىــ, گۈلىــ>,“ئالىم(阿里木)”和“ئۇنىــ(他)”构成了正例<ئۇنىــ, ئالىم>。

根据上述的十一个特征,提取的特征向量值如表 1 所示。

表 1　训练和测试实例格式

| <照应语,先行语> | $V_{ap}$ | $V_{cp}$ | $V_{nest}$ | $V_{ga}$ | $V_{sa}$ | $V_{na}$ | $V_{pos}$ | $V_{name}$ | $V_{role}$ | $V_{cg}$ | $V_{distance}$ | tag |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| <ئۇنىــ, ئالىم> | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| <ئۇنىــ, گۈلىــ> | 1 | 0 | 0 | 0 | 0.5 | 1 | 0 | 1 | 0 | 0 | 1 | -1 |

## 4　实验结果与分析

### 4.1　语料来源

实验语料来自天山网等维吾尔语网页网站。首先用网络爬虫在网上下载网页,然后经过去重和降噪后筛选出包含小说等内容作为实验语料。在维吾尔语语言学专家的帮助指导下,标注完成的语料共 300 篇。实验语料中第一、二和三人称代词占比分别为:35.36%、11.42%、53.23%。

### 4.2　实验测评标准

利用自然语言处理中常用的 MUC 标准对实验结果进行测评。准确率 $P$:模型的准确程度;召回率 $R$:模型的完备性;$F_1$ 值:指代消解性能,表达式为:

$$P = \frac{\text{正确消解的实例数目}}{\text{模型识别的实例数目}} \times 100\% \tag{10}$$

$$R = \frac{\text{正确消解的实例数目}}{\text{总的消解实例数目}} \times 100\% \tag{11}$$

$$F_1 = \frac{2 \times R \times R}{R + R} \times 100\% \tag{12}$$

### 4.3　实验设计

为了确保实验结果的有效性,避免实验的不确定性,在进行实验时,将实验样本全部随机打乱,确保数据的随机性。实验采用五倍交叉验证,取其平均值作为实验结果。参数设置如下:学习率 0.01;批处理样本数 15;词向量维度 150;迭代次数 100;LSTM 隐藏层节点数目 110;RBM 层数 2。

### 4.4　Word Embedding 对实验的影响

文中采用 Word Embedding 将词汇向量化表示作为本文模型输入的数据。Word Embedding 区别于传统的文本数据表示方法,提供了更好的语义特征信息,可以避免传统词向量的维度过高的问题,并且解决了向量稀疏问题,从而降低了模型的训练难度。

WordEmbedding 的不同维度,对指代消解的性能也有一定的影响,维度越高含有的语义信息也越多。为了探索不同维度的词向量对实验结果的影响,文中分别将 10 维、50 维、100 维、150 维、200 维的词向量作为模型的输入数据。实验结果如表 2 所示。

由表 2 可知,Word Embedding 的维度选择对模型的准确率有很大的影响。随着 Word Embedding 维度的增加,反映整体性能的 $F_1$ 值也逐步提高,并在 Word Embedding 维度达到 150 维时,综合值 $F_1$、准确率 $P$ 和召回率 $R$ 均达到了最高值,使实验获得了最优的效果,$F_1$ 值也达到了 78.83%,准确率达到了 81.14%。当将 Word Embedding 的维度继续增加时,综合值 $F_1$ 却没有继续增加,反而降低了;这是因为高维度向量中虽然包含了丰富的语义信息,但是也引入了噪音和无用的干扰信息,会产生过拟合现象,造成模型对数据的泛化能力降低,影响了模型指代消解的性能。

表 2　不同维度下指代消解性能对比　%

| 维度 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 10 | 68.52 | 63.60 | 65.97 |
| 50 | 72.82 | 68.54 | 70.62 |
| 100 | 77.31 | 70.25 | 73.61 |
| 150 | 81.14 | 76.65 | 78.83 |
| 200 | 71.24 | 69.40 | 70.31 |

### 4.5　模型对比实验

为了验证模型的有效性,将文中模型与传统 LSTM、LSTM、DBN 等深度学习模型进行对比,结果如表 3 所示。

表 3　模型对比结果　%

| 模型 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 文中模型 | 81.14 | 76.65 | 78.83 |
| DBN | 75.99 | 74.42 | 75.20 |
| 传统 LSTM | 74.11 | 71.32 | 72.69 |
| LSTM | 77.18 | 73.45 | 75.27 |

从表 3 可知,LSTM 模型在准确率、召回率、综合值等指标上均高于传统的 LSTM 模型,这是因为 LSTM 模型充分利用了短时信息记忆功能的记忆单元,能够将上一时刻存储的关键词汇信息用于挖掘下一时刻的词汇语义信息。文中模型比 LSTM 实验性能更优,是因为当输入文本过长时,LSTM 模型不仅容易丢失大量的细节信息,且不能很好地分配权重比,造成信息的缺失,从而影响模型的性能。因此文中模型加入了 Attention 机制。注意力机制能有效降低数据维度、提高计算速度,将输入的长文本映射成含有语义信息的数据编码,避免造成信息的缺失。单一的 DBN 模型在其评价标准上比 Attention-Based LSTM-DBN 模型的相对较低,是因为 Attention-Based LSTM-DBN 模型中,

长短时记忆网络模型能够更好地联系上下文,挖掘出人称代词语义信息,受限玻尔兹曼机网络能够保证特征向量达到最优化,挖掘出更深层次的语义特征,从而提高输出质量。结果表明,文中模型在维吾尔语人称代词指代消解研究中性能够优。

### 4.6 与其他模型对比实验

在同等条件下,将文中模型与 SVM、SAE、ANN 进行对比,结果如表 4 所示。

表 4 与其他模型实验对比结果 %

| 模型 | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| 文中模型 | 81.14 | 76.65 | 78.83 |
| SVM | 67.15 | 74.19 | 70.49 |
| ANN | 70.93 | 71.77 | 71.34 |

由表 4 可知,3 种模型中,SVM 和 ANN 在准确率、召回率、综合值均低于文中模型。这是因为浅层机器学习模型 SVM 和 ANN,相较于 Attention-Based LSTM-DBN 挖掘文本数据中隐藏的深层语义信息的能力相对较差,不能更好地利用数据中隐藏的信息。而文中利用深层神经网络构建的人称代词指代消解模型,能够更好地适应复杂的数据分布情况,挖掘出更深层次的语义信息。因此文中模型相较于浅层机器学习更适用于代词的消解研究。

## 5 结束语

维吾尔语人称代词指代消解对于维吾尔语自然语言领域的研究和发展具有重要的意义。目前在自然语言领域的研究主要针对的是英语、汉语等大语种,而针对维吾尔语等小语种的指代消解的研究相对较少,此外也没有充分考虑上下文的语义信息,数据转换过程中信息丢失严重,不能够很好地挖掘出更深层次的语义特征。针对这些问题,采用 Attention-Based LSTM-DBN 模型,对文章的上下文语义特征进行挖掘。并且利用词向量将文本转换成含有丰富语义信息的特征向量作为模型的输入。根据维吾尔语人称代词指代的现象抽取 11 项规则特征,利用两类融合后的特征,完成维吾尔语人称代词指代消解研究。通过与长短时记忆网络等模型进行对比实验,验证了该模型在篇章级别文本上挖掘深层语义特征的有效性,提高了维语人称代词指代消解的性能。而与其他模型进行的对比实验,验证了 Attention-Based LSTM-DBN 模型在挖掘深层次的维吾尔语人称代词语义信息方面比浅层机器学习算法更具优势,能更好地应对复杂的数据分布情况。

**参考文献:**


[1] 奚雪峰,周国栋. 基于 Deep Learning 的代词指代消解[J]. 北京大学学报:自然科学版,2014,50(1):100-110.

[2] MCCARTHYJ F, LEHNERT W G. Using decision trees for conference resolution[C]//Proceedings of the 14th international joint conference on artificial intelligence. Montréal Québec, Canada: Morgan Kaufmann Publishers Inc., 1995: 1050-1055.

[3] 王荣波,孙小雪,黄孝喜,等. 基于指代消解的汉语句群自动划分方法[J]. 计算机技术与发展,2017,27(8):61-65.

[4] 罗云飞,李国臣. 采用优先选择策略的中文人称代词的指代消解[J]. 中文信息学报,2005,19(4):24-30.

[5] NG V. Semantic class induction and conference resolution [C]//Proceedings of the meeting of the association for computational linguistics. [s. l. ]:[s. n. ],2007:536-543.

[6] KONG Fang, ZHOU Guodong, ZHU Qiaoming. Employing the centering theory in pronoun resolution from the semantic perspective[C]//Proceedings of the 2009 conference on empirical methods in natural language processing. Singapore: ACL, 2009:987-996.

[7] 许 敏,王能忠,马彦华. 汉语中指代问题的研究及讨论[J]. 西南师范大学学报:自然科学版,1999,24(6):633-637.

[8] 王厚峰,何婷婷. 汉语中人称代词的消解研究[J]. 计算机学报,2001,24(2):136-143.

[9] 王厚峰,梅 铮. 鲁棒性的汉语人称代词消解[J]. 软件学报,2005,16(5):700-707.

[10] 董国志,朱玉全,程显毅. 中文人称代词指代消解的研究[J]. 计算机应用研究,2011,28(5):1774-1776.

[11] 王海东,胡乃全,孔 芳,等. 基于树核函数的英文代词消解研究[J]. 中文信息学报,2009,23(5):33-39.

[12] 孔 芳,周国栋. 基于树核函数的中英文代词消解[J]. 软件学报,2012,34(5):1085-1099.

[13] 孙茂松,刘 挺,姬东鸿,等. 语言计算的重要国际前沿[J]. 中文信息学报,2014,28(1):1-8.

[14] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]//International conference on machine learning. Helsinki, Finland: ACM, 2008:160-167.

[15] 胡乃全. 基于特征向量的中文指代消解研究与系统实现[D]. 苏州:苏州大学,2009.

[16] MNIH A, HINTON G. Three new graphical models for statistical language modelling[C]//International conference on machine learning. Corvalis, Oregon, USA: ACM, 2007:641-648.

[17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[18] 胡新辰. 基于 LSTM 的语义关系分类研究[D]. 哈尔滨:哈尔滨工业大学,2015.

[19] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Conference on empirical methods in natural language processing. Lisbon, Portugal: ACL, 2015:1412-1421.

[20] 张 冲. 基于 Attention-Based LSTM 模型的文本分类技术的研究[D]. 南京:南京大学,2016.