

文章编号: 1003-0077(2019)09-0060-09

基于 ATT-IndRNN-CNN 的维吾尔语名词指代消解

祁青山¹, 田生伟¹, 禹 龙², 艾山·吾买尔²

(1. 新疆大学 软件学院, 新疆 乌鲁木齐 830091;
2. 新疆大学 信息科学与工程学院, 新疆 乌鲁木齐 830046)

摘 要: 该文提出一种基于注意力机制(attention mechanism, ATT)、独立循环神经网络(independently recurrent neural network, IndRNN)和卷积神经网络(convolutional neural network, CNN)结合的维吾尔语名词指代消解模型(ATT-IndRNN-CNN)。根据维吾尔语的语法和语义结构,提取 17 种规则和语义信息特征。利用注意力机制作为模型特征的选择组件计算特征与消解结果的关联度,结果分别输入 IndRNN 和 CNN 得到包含上下文信息的全局特征和局部特征,最后融合两类特征并使用 softmax 进行分类完成消解任务。实验结果表明,该方法优于传统模型,准确率为 87.23%,召回率为 88.80%,F 值为 88.04%,由此证明了该模型的有效性。

关键词: 注意力机制;独立循环神经网络;CNN;指代消解;维吾尔语

中图分类号: TP391 **文献标识码:** A

Anaphora Resolution of Uyghur Nouns Based on ATT-IndRNN-CNN

QI Qingshan¹, TIAN Shengwei¹, YU Long², AISHAN Wumaier²

(1. School of Software, Xinjiang University, Urumqi, Xinjiang 830091, China;
2. College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, China)

Abstract: This paper proposes an Uyghur nouns anaphora resolution model ATT-IndRNN-CNN based on Attention Mechanism (ATT), Independently Recurrent Neural Network (IndRNN) and Convolutional Neural Network (CNN). According to the grammar and semantic structure of Uyghur, 17 rules and semantic information features are extracted. The attention mechanism is applied to select the features via the correlation between the features and the resolution results. The results are input into IndRNN and CNN to obtain the global features and local features in the context, respectively. Finally, the two types of features are merged and softmax is used to classify the resolution task. The experimental results show that the proposed method is better than the classical models, achieving the precision of 87.23%, the recall of 88.80%, and the F-measure of 88.04%.

Keywords: attention mechanism; IndRNN; CNN; anaphora resolution; Uyghur

0 引言

指代(anaphora)是在自然语言中常见的一种语言现象,在篇章中通常利用一个抽象的词语代替前面的某个具体的词语。语言学中将抽象的语言单位称为照应语(anaphor),而具体的实体称为先行语(antecedent)。确定某个照应语的先行语的过程称为指代消解^[1]。指代消解对于自然语言处理(natu-

ral language processing, NLP)研究中的机器翻译(machine translation)、信息抽取(information extraction)、自动文摘(automatic abstracting)以及自动问答(question answering)等自然语言应用系统都具有非常重要的支撑作用^[2]。指代消解分为指代(anaphora)和共指(coreference),回指也称为指示性指代,是当前的词语与上文中出现的具体的词语具有密切联系;共指也称为同指,是两个具体的词语对应于现实世界中共同参照物的指代^[3]。

指代消解发展的几十年来,已经从基本的基于规则的研究方法逐渐过渡到机器学习的研究方法中。McCarthy^[4]等首先提出将指代消解改成二分分类问题,在候选先行语中判断与照应语是否具有紧密的联系,进而判断是否具有指代关系。Soon^[5]等在此基础上提出使用机器学习进行指代消解的研究框架并给出了可用的系统。他们在语料中提取 12 种特征作为分类的标准,再利用支持向量机对其进行训练得到分类模型。这一思想影响了后期大多数学者的研究。Ng^[6]等在 Soon 的基础上将特征扩充成 53 种;Yang^[7]等提出了一种双候选模型,可以更好地确定照应语对应的先行语;Kong^[8]等把中心理论应用到语义层,提高了指代消解的性能。在中文指代消解中,郭志立^[9]提出了利用人称代词本身的语义信息等进行人称代词先行语的分析;马彦华^[10]等采用了一种“主题人物法”的方法来解决中文的人称代词消解问题;许敏^[11]等利用上下文中的语义信息进行指代分类;王厚峰^[12-14]等对汉语的指代消解有较多的研究。但是对于像维吾尔语这种小语种的研究较少,主要有李冬白^[15]等提出了一种基于 DBN 深度神经网络学习模型的方法对维吾尔语的人称代词进行消解;李敏^[16]等提出了一种基于栈式自编码深度学习的维吾尔语名词消解方法。

随着研究不断深入,研究者们发现对于特征给予不同的关注程度可以更好地进行分类,并且在篇章中上下文信息对于指代消解也具有极其重要的作用。此外,目前的研究大多都集中于中文和英文,针对维吾尔语这种语料资源匮乏的小语种的研究非常少,并且将深度学习用到维吾尔语的名词指代消解中的研究也很少。基于上述问题,本文提出一种基于注意力机制、独立循环神经网络和卷积神经网络相组合的方法,用于维吾尔语的名词指代消解。在该方法中先利用注意力机制作为模型特征的选择组件计算特征的权重,使得特征与消解结果的联系更加紧密。再利用独立循环网络和卷积神经网络分别得到全局特征和局部特征,并将这两种特征进行融合,在得到上下文信息的同时又不丢失局部信息,可以得到更好的分类结果,提升维吾尔语名词指代消解的性能。

1 相关知识

1.1 指代消解

指代消解是自然语言中的一个语言单位用于确

定其指向之前出现的语言单位的过程。其中用于指向的语言单位,称为照应语(anaphors),被指向的语言单位称为先行语(antecedent)。根据消息理解会议(Message Understanding Conference, MUC)对指代的定义,认为指代关系不仅仅存在于代词与名词(名词短语)之间,还存在于名词(名词短语)与名词(名词短语)之间。例如,

例 1

ئاخىر بىر كۈنى مەن ئانامنىڭ غەلىتە بىر پارچە خېتىنى تاپشۇرۇپ ئالدىم، بۇ خەتنى ئاچام ئاينۇرغا ئەۋەتتىلەتتى. ئانامنىڭ ئۇنىڭغا يازغان بۇ بىر پارچە خېتىنى تولۇق ئوقۇش ئۈچۈن مەن چوقۇم ئاچام بىلەن بىر قېتىم كۆرۈشۈشمىگە توغرا كېلىمىتى

(最终有一天,我收到了母亲一封奇怪的信,这封信是寄给姐姐阿依努尔的。为了读母亲给她写的这一封完整的信,就需要我必须跟姐姐见一次面)

例 1 选自实验语料,其中存在很多的指代关系,包括名词与代词之间的指代,“ئاچام(姐姐)”和“ئۇ(她)”;名词短语与代词之间的指代:“ئاينۇرغا ئاچام(姐姐阿依努尔)”和“ئۇ(她)”;名词短语与名词之间的指代:“خېتىنى غەلىتە بىر پارچە(奇怪的一封信)”和“خەتنى(信)”;名词短语与名词短语之间的指代:“غەلىتە بىر پارچە خېتىنى(奇怪的一封信)”和“بۇ بىر پارچە خېتىنى تولۇق(这一封完整的信)”。在进行指代消解时,对实体以及对应指代信息的高效识别可以提高指代消解系统的性能。

1.2 维吾尔语的特点

维吾尔语是一种带格语法的黏着性语言,词组与句子之间有严格的词序,并且拥有“格”结构。这种“格”结构对于指代消解工作起到非常重要的作用,利用维吾尔语的格语法可以判断词语的词性等重要内容。在维吾尔语中一般认为有 6 种格,具体如表 1 所示。

表 1 维吾尔语的格

格范畴	含义
主格	表示句子中的主语,没有后缀
属格	表示句子中人或者事物的所属关系,后缀为:ئىك
位格	在句子中表示时间、地点、环境等描述的状态空间,后缀为:تە، دە، تا، دا
向格	表示对象、目的、方式等针对某一动作或者性质的事物,后缀为:قا، غا، كە، گە
宾格	表示动作的对象,后缀为:نى

续表	
格范畴	含义
从格	表示起点、经过的处所、时间段、间接宾语和比较的成分,后缀为: تىن, دىن

维吾尔语中的名词存在单复数变化,可以将名词单复数作为一个非常重要的特征,判断是否具有指代关系,排除不存在指代关系的样本,这也为维吾尔语名词的指代消解提供了较好的基础。

2 模型介绍

针对维吾尔语名词指代消解问题,本文利用 Soon^[5]等提出的框架,首先确定照应语的候选先行语,提取名词短语的特征,再引入注意力机制(attention)赋予特征权重,将得到的带权特征分别输入到独立循环神经网络(IndRNN)模型和卷积神经网络,得到包含上下文信息的全局特征和局部特征,最后将得到的全局特征和局部特征进行融合,放入 Softmax 中训练分类,模型如图 1 所示。

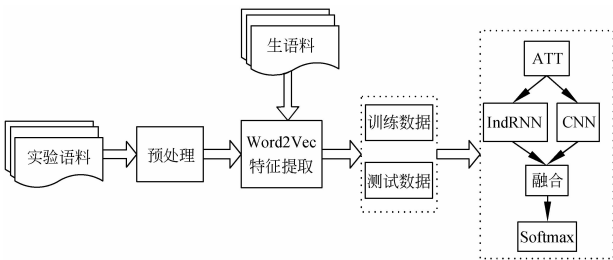


图 1 维吾尔语名词短语消解框架

2.1 特征提取

在自然语言处理中特征提取是一项非常重要的工作,提取的特征是否具有代表性和通用性直接决定了最后实验结果的好坏。而特征提取在指代消解中起到的作用更大,因此本文结合实验组维吾尔语言学专家总结的具有指称性的名词短语以及前人的经验选取以下特征进行指代消解。

2.1.1 规则特征

规则特征是根据语言内部结构规则进行提取的特征,主要体现先行语和照应语在文章内部的关系,本文主要提取了 17 种规则特征,具体如表 2 所示。

表 2 规则特征

特征名称	描述对象	特征描述
z_pronoun	照应语	若照应语为人称代词,如 1.1 节例 1 中的 ئۇ(她),该特征值为 1;若照应语为反身代词,如 ئۆزىنىڭ(自己的),该特征值为 2;若非代词,则取 0
x_pronoun	先行语	若先行语为人称代词,该特征值为 1;若先行语为反身代词,该特征值为 2;若非代词则取 0
z_proper_noun	照应语	若照应语是专有名词,该特征值为 1,否则为 0。专有名词包括人名、地名、节日等名词,如例 1 中的人名 ئاينۇر(阿依努尔)
x_proper_noun	先行语	若先行语是专有名词,该特征值为 1,否则为 0
z_concrete_noun	照应语	若照应语为具体名词,该特征值为 1,否则为 0。具体名词指的是某些具体事物,而非抽象的属性等名词,如例 1 中的具体名词 خېتىنى(信)
x_concrete_noun	先行语	若先行语为具体名词,该特征值为 1,否则为 0
z_attribute_noun	照应语	若照应语是有如形容词、代词、数词等定语修饰的名词,则该特征值取 1,否则为 0。如例 1 中 غەلىتە بىر پارچە خېتىنى(奇怪的一封信)中的名词 خېتىنى(信)为定语修饰的名词
z_indicative_noun	照应语	若照应语为指示性名词,该特征值为 1,否则为 0
z_formate_noun	照应语	若照应语为主格结构,则该特征值为 1;若为宾格结构,该特征值为 2;其他为 0
x_formate_noun	先行语	若先行语为主格结构,则该特征值为 1;若为宾格结构,该特征值为 2;其他为 0。如例 1 中的 خېتىنى(信)是 ئەۋەتلىپتۇ(寄)的动作对象,所以为宾格结构,故特征应设置为 2
part_of_speech	两者关系	若先行语和照应语的词性一致,则该特征值为 1,否则为 0
singular_or_plural	两者关系	若先行语和照应语满足单复数一致原则,该特征值为 1,否则为 0。如例 1 中的先行语 ئاينۇرغا ئاچام(姐姐阿依努尔)和照应语 ئۇ(她)都是单数,所以该特征值为 1

续表

特征名称	描述对象	特征描述
distance	两者关系	该特征描述先行语和照应语的距离关系,具体的计算方法为: $d = \begin{cases} 0 & i > 10 \\ 1 - (0.1 * i) & 0 \leq i \leq 9 \end{cases}$ 其中, i 为先行语和照应语相距的句子的数目,当先行语和照应语在同一句时 i 取 0
semantic_category	两者关系	若照应语与先行语的语义类别相同,则该特征为 1,否则为 0。实验语料中的语义类别是通过维吾尔语学专家进行标注的,本文标注了 14 种语义类别,如例 1 中的先行语 ئاينۇرغا ئاچام (姐姐阿依努尔)和照应语 ئۇ(她)的语义类别都是人类,故特征值应为 1
sex	两者关系	若照应语和先行语的性别一致,则该特征值为 1;若其中一个未知,则该特征值为 0.5;若不同,则该特征值为 0
appositive	两者关系	若照应语和先行语是同位语关系,则该特征取 1,否则取 0
key_word	两者关系	若照应语和先行语满足中心词匹配,该特征取 1,否则取 0。其中,中心词是事先由维吾尔语专家进行标注的

2.1.2 语义特征

在指代消解工作中,提取规则特征虽然可以进行消解工作,但是缺少对整个句子语义的考虑。因此本文采用词向量的方式将先行语和照应语在句中深层次的语义特征表现出来。为了避免维度灾难,本文采取了 Mikolov 等^[17]在 2013 年提出的 Word2Vec 工具进行词向量的训练。同时为了准确得到词语在多维空间中的语义分布情况,对原有语料进行了扩充。利用爬虫从人民网和天山网等网站的维吾尔语板块爬取维吾尔语文本,并进行简单降噪处理,得到 8 000 篇未标注的生语料文本,经过分词处理后得到 1 003 267 个分词数据,与实验语料进行结合,训练照应语和先行语的语义特征。

2.2 训练和测试样本构成

本文将语料文本进行预处理,再经过维吾尔语言学专家对语料库进行词性和相应的指代链标注。通过对进行标注的照应语在句子中出现的位置提取上下文的名词构成候选先行语集合,再将其遍历,判断是否是该照应语的先行语,若是则形成正例样本,否则形成负例样本。具体算法为:

Step1 提取单个文本中所有的名词短语,根据标注判断是否为照应语,若是,则存入集合 {anaphors},否则存入集合 {nouns} 中。

Step2 遍历集合 {anaphors},将每一个照应语 anaphor 与集合 {nouns} 中每个元素 noun 进行对比,若两个元素属于同一指代链,则将标签标记为 1,作为正例;否则将标签标记为 0,作为负例。同时根据 2.1.1 节中表 2 中所提到的特征,读取文本中这两个元素的信息,进行对比,构成样本。

Step3 重复 step1、step2,直到将所有的语料遍历一遍。

通过上述算法得到全部样本,并且将其中的 80%作为训练数据集,20%作为测试数据集。

2.3 ATT-IndRNN-CNN

ATT-IndRNN-CNN 模型结合注意力机制与两种不同的神经网络,可以有效地将全局特征与局部特征进行组合,该模型对数据处理主要分为 3 个阶段:首先利用注意力机制强化特征,然后将处理后的特征分别输入 IndRNN 和 CNN 得到全局特征和局部特征,最后将全局特征和局部特征进行融合,形成新的特征,输入 Softmax 进行分类训练。模型总框架如图 2 所示。

2.3.1 注意力机制(ATT)

注意力机制最早是在图像处理中的被提出来的,Mnih^[18]等将之用于图像分类的同时 Bahdanau^[19]也将其用到了机器翻译之中。本文中利用注意力机制主要是将不同的特征赋予不同的权重,以便更好地进行模型的训练。在注意力机制中将输入特征 Data 看做由<Key,Value>数据对组成,对于给定元素 Q 的 Attention 值计算如式(1)所示。

$$\text{attention}(Q, \text{Data}) = \sum_{j=1}^{L_a} a_j \cdot \text{Value}_j \quad (1)$$

其中, L_a 为 Data 的长度, a_j 为 value_j 对应的权重系数, a_j 求解方式如式(2)所示。

$$a_j = \text{softmax}(\text{sim}_j) = \frac{e^{\text{sim}_j}}{\sum_{k=1}^{L_a} e^{\text{sim}_k}} \quad (2)$$

利用 Softmax 对 Q 和各个 Key 之间相似度数值进行归一化,同时也利用 Softmax 内在的机制突

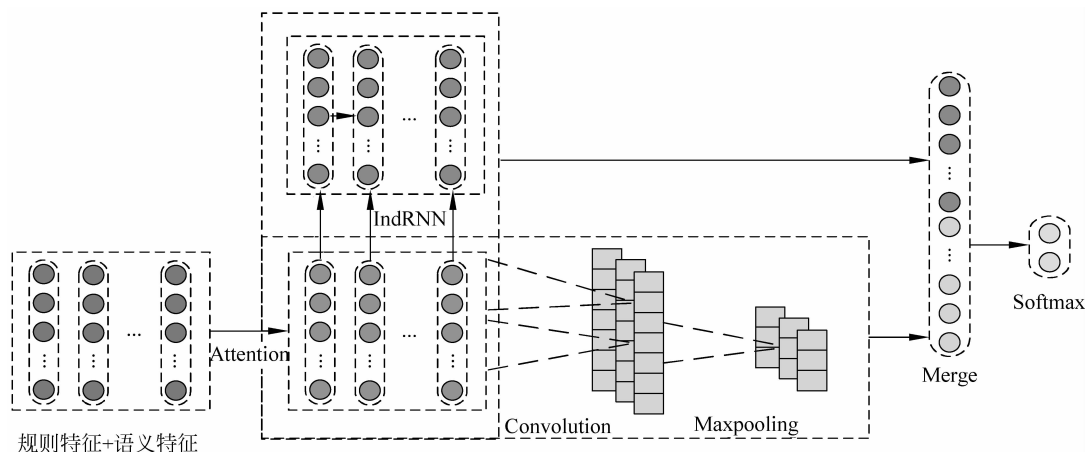


图 2 ATT-IndRNN-CNN 模型

出重要成分的权重,而 Q 和各个 Key 之间相似度计算通常通过计算两者之间的点积来得到,如式(3)所示。

$$\text{sim}(Q, \text{Key}_j) = Q \cdot \text{Key}_j \tag{3}$$

通过上面一系列的计算可以得到特定元素 Q 的 Attention 数值。将得到特征经过注意力机制可以突出其中某些特征权重信息,进而可以更加有效地进行分析。

2.3.2 独立循环神经网络(IndRNN)

独立循环神经网络是由 Li 等人提出的一种新型循环神经网络^[20],这种新型循环神经网络可以有效地解决普通 RNN 在训练收敛时存在的梯度爆炸和梯度消失的问题,同时可以处理更长的序列。其基本计算如式(4)所示。

$$h_t = \sigma(Wx_t + u \odot h_{t-1} + b) \tag{4}$$

其中, σ 为神经元的逐元素激活函数, u 为一个循环权重向量, W 为当前权重, b 为神经元偏差, \odot 表示 u 与 h_{t-1} 的阿达马积。基本结构图如图 3 所示。

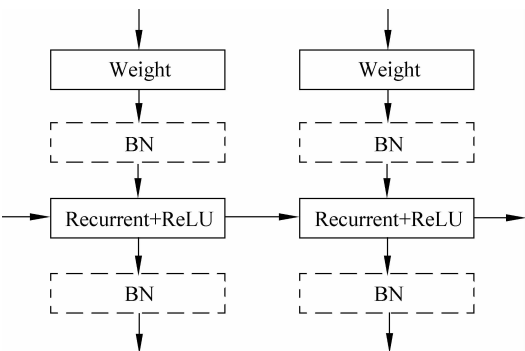


图 3 IndRNN 结构

IndRNN 中每层的神经元是相互独立的,但是

可以将 IndRNN 进行多层叠加,并且层与层之间的神经元进行连接。对于隐含层第 n 个神经元的 $h_{n,t}$ 可以通过式(5)进行计算。

$$h_{n,t} = \sigma(\omega_n x_t + \mu_n h_{n,t-1} + b_n) \tag{5}$$

其中, μ_n 表示第 n 行的循环输入权重,而 ω_n 表示第 n 行的当前输入权重, b_n 为第 n 行的神经元偏差。由式(5)可以看出,每个神经元仅接收当前状态隐藏层和输入其中的信息,各个神经元之间都是相互独立的时空特征,这就使得 IndRNN 可以方便地进行组合。本文将经过 ATT 处理的特征输入到两层的 IndRNN 中,每层 IndRNN 包含 64 个隐含单元,得到包含上下文信息的全局特征。

2.3.3 卷积神经网络(CNN)

CNN 是一种前馈神经网络,前期主要应用在图形处理中,可以避免前期复杂的图像处理。近年来研究者们将 CNN 引入自然语言处理可以有效缓解特征工程中的工作量,并且可以得到局部特征。CNN 如图 4 所示,主要包括输入层、卷积层、池化层。

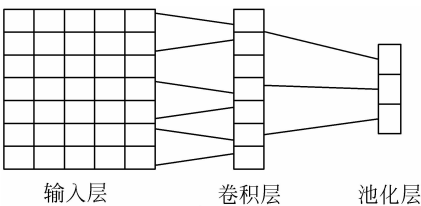


图 4 卷积神经网络

输入层为 ATT 的输出特征,经过卷积层利用卷积核对局特征进行卷积处理得到局部更具代表性的特征。基本计算如式(6)所示。

$$c = f(W \cdot x + b) \tag{6}$$

其中, x 为卷积核窗口词向量矩阵, W 为权重

矩阵, b 为偏置, f 为激活函数。池化层是卷积神经网络的重要网络层, 该层可以对卷积层得到的特征向量进行采样, 进一步调整卷积层的输出。池化函数利用某一位置的相邻输出的总体统计特征来代替网络在该位置的输出。当我们重点关注某个特征是否出现而不是出现的具体位置时就要利用到局部平移不变性, 而池化就实现了这一点。一般的池化函数有最大池化函数和平均池化函数之分, 本文使用最大池化函数。

2.3.4 特征融合

这一阶段将 2.3.2 和 2.3.3 得到的特征进行融合, 本文使用张量相乘的方法对两种特征进行连接, 对于两个特征 V 和 U 其张量乘积 $V \otimes U$ 计算定义如式(7)所示。

$$(V \otimes U)_{i_1 i_2 \dots i_{m+n}} = V_{i_1 i_2 i_3 \dots i_n} U_{i_{n+1} i_{n+2} \dots i_{n+m}} \quad (7)$$

其中, n 和 m 分别为 V 和 U 的协变张量。利用张量积可以将两个张量融合, 并且张量积继承了其因子的所有指标, 不丢失原本张量的信息。

3 实验和分析

3.1 语料准备

基于机器学习的指代消解的方法是需要相应的语料支撑的, 目前进行的英文的指代消解的语料常用消息理解会议(Message Understanding Conference, MUC), 中文指代消解采用的语料大多数是自动内容抽取会议(Automatic Content Extaction, ACE)或者 OntoNotes 的语料, 但是目前关于维吾尔语的已标注的语料尚未见公开报道, 因此需要针对维吾尔语名词指代消解对维吾尔语语料进行筛选和标注。

本文利用网络爬虫从人民网和天山网等网站的维吾尔语板块爬取的文章中筛选出存在指代链信息, 在维吾尔语专家的指导下对其进行标注, 包括标注指代链信息、名词短语、语义类别、名词单复数、格语法等特征, 对标注后的语料利用 Excel 文件进行存储。

本实验中共标注了 370 篇文章, 其中包含 19 553 条实体名词, 9 725 条动词和 3 239 条代词, 标注的语法结构包括 6 172 条主语, 8 232 条谓语, 6 518 条宾语, 6 984 条定语和 10 335 条状语。语料中共有 17 046 条词语包含语义类别, 13 265 条词语拥有格属性。利用 2.2 节中提出的方法形成训练和

测试数据集共 75 084 组数据, 其中包括具有指代关系的 20 266 组正例和不具有指代关系的 54 818 组负例。

3.2 实验结果与分析

为了方便实验结果的对比, 本文采用自然语言处理经常采用的 3 种测评标准: 准确率 P 、召回率 R 和 F 值, 对实验进行测评。其中 P 可以反映模型的准确率, R 可以反映模型查全率, F 值可以很好地综合考虑 P 和 R 进而反映模型的综合性能, F 值的计算如式(8)所示。

$$F = \frac{R \times P \times 2}{R + P} \quad (8)$$

同时, 为了实验结果的稳定性和代表性, 本实验采用 5 折交叉验证, 取平均值作为最终实验结果。每次实验均利用 GPU GTX 1050 提高运行速率, 进而减少运行时间。本文对不同的参数组合进行了反复实验, 确定实验中各个模型的最优参数。后续实验均采用最优参数进行实验。最优参数如表 3 所示。

表 3 参数设置

参数	参数取值
ϵ	0.001
batch	32
act	relu
filters	64
filter-size	3
dropout	0.2
epochs	500

其中, ϵ 表示训练过程中的学习率, batch 表示每次迭代时批量处理的个数, act 表示模型的激活函数, filters 表示 CNN 中卷积核的数目, filter-size 表示 CNN 中卷积核的大小, dropout 表示在训练过程中的丢码率, epochs 表示迭代的次数。迭代次数对实验结果的影响如图 5 所示。

3.2.1 ATT-IndRNN-CNN 模型的有效性

在维吾尔语名词指代消解上, 为了验证本文提出的模型的有效性, 在特征相同的情况下, 对不同的神经网络进行了对比实验, 为了使得实验更具有说服力, 各个网络均在自己的最优参数下进行实验, 实验结果如表 4 所示。

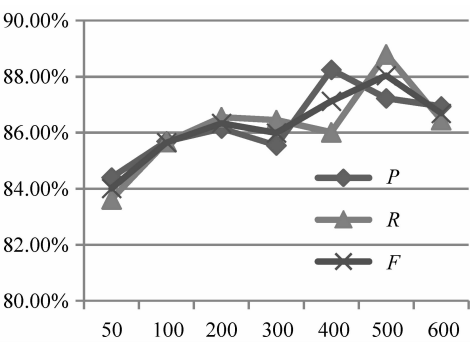


图 5 迭代次数对实验结果的影响

表 4 模型对比

模型	P / %	R / %	F / %
CNN	84.36	83.17	83.70
IndRNN	82.26	77.48	81.63
ATT-CNN	84.67	84.9	84.79
ATT-IndRNN	85.92	87.13	86.52
ATT-IndRNN-CNN	87.23	88.80	88.04

表 4 表明在单独的神经网络中加入注意力机制时,ATT-CNN 比 CNN 的 F 值提高了 1.09%,ATT-IndRNN 比单独 IndRNN 的 F 值提高了 4.89%。这说明当在模型中加入注意力机制时,可以使模型的性能在一定程度上有所提高。当使用本文提出的 ATT-IndRNN-CNN 联合模型方法时,准确率 P 、召回率 R 和 F 值较之单一模型或者加入注意力机制的单一模型均有提高,充分说明了本文方法的有效性。

3.2.2 语义特征对指代消解的影响

2.1.1 节的规则特征仅考虑了先行语和照应语之间的关系,对两者在句子中的语义内容考虑得较少,因此本节针对基于词向量模式的语义特征对指代消解的影响进行了对比实验,实验在原有规则特征的基础上引入 100 维的词向量语义特征,分别对不同的模型进行对比实验,并且对实验耗时进行了记录。实验结果如图 6 所示,耗时结果如表 5 所示(表 5 中 CNN+W 表示在原有规则特征基础上添加语义特征,其他类似)。

表 5 模型耗时对比

模型	耗时
CNN	40 min 49 s
CNN+W	44 min 43 s
IndRNN	1 h 16 min 19 s

续表

模型	耗时
IndRNN+W	1 h 17 min 48 s
ATT-CNN	54 min 52 s
ATT-CNN+W	56 min 48 s
ATT-IndRNN	1 h 22 min 34 s
ATT-IndRNN+W	1 h 20 min 21 s
ATT-IndRNN-CNN	1h 44 min 12 s
ATT-IndRNN-CNN+W	1h 47 min 18s

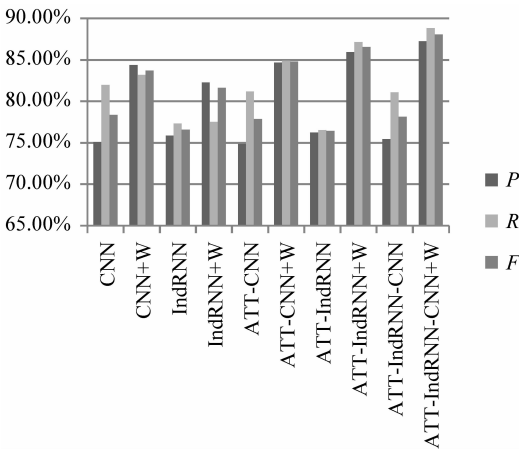


图 6 语义特征对比

由图 6 可以看出,在加入语义特征后所有模型的准确率 P 、召回率 R 和 F 值均有显著提高,实验结果充分说明了加入语义特征向量的有效性,这是因为规则特征仅包含先行语和照应语之间的结构特点,缺乏对整个句子中语义信息的考虑,而加入词向量融合特征后可以对先行语和照应语在句子中的语义信息进行建模,进而提高指代消解的准确性。

由表 5 可以看出,在增加语义特征时,模型耗时有少量的增加,说明加入语义特征后对于模型耗时影响较小。但将模型进行融合后耗时明显变长,表明运行时间受模型影响较大。

3.2.3 词向量维度对指代消解的影响

在融合语义特征时,训练的词向量维度的大小也会影响实验结果,理论上向量的维度越高,包含的语义信息也就会越丰富,因此本文分别采用 10 维、30 维、60 维、100 维和 150 维的词向量进行了对比实验,实验结果如表 6 所示。

由表 6 可以明显地看出,在词向量维度达到 100 时,准确率 P 、召回率 R 和 F 值都达到最优,而当维度为 150 时,性能有所下降。这是因为当维度

表 6 词向量维度对比

维度	<i>P</i> /%	<i>R</i> /%	<i>F</i> /%
10	83.39	86.29	84.81
30	86.05	85.05	85.56
60	86.45	84.12	85.29
100	87.23	88.8	88.04
150	86.91	84.93	85.91

越高时,包含的信息越多,就越有可能产生过拟合的现象,从而导致模型对数据的泛化能力降低。

3.2.4 规则特征对实验的影响

为了证明本文提取的人工特征对实验的影响,本文进行了以下的对比实验。为了使一个规则特征可以得到有效训练,因此加入 10 维的词语向量,逐渐增加人工特征数目,其他设置按照 3.2 节中表 3 的最优设置。采用本文提出的模型进行实验得到准确率 *P*、召回率 *R* 和 *F* 值如表 7 所示。

表 7 规则特征对结果的影响

特征	<i>P</i> /%	<i>R</i> /%	<i>F</i> /%
z_pronoun	76.72	55.2	64.21
x_pronoun	75.65	56.85	64.91
z_proper_noun	79.78	60.82	69.02
x_proper_noun	86	58.93	69.94
z_concrete_noun	77.99	82.03	79.96
x_concrete_noun	79.77	80.6	80.18
z_attribute_noun	81.56	79.13	80.33
z_indicative_noun	80.46	80.34	80.41
z_formate_noun	81.52	79.7	80.6
x_formate_noun	82.58	79.17	80.83
part_of_speech	80.7	81.81	81.25
singular_or_plural	85.76	78.19	81.8
distance	82.48	81.32	81.9
semantic_category	81.4	82.76	82.07
sex	83.25	81.32	82.27
appositive	81.46	85.44	83.41
key_word	83.39	86.29	84.81

由表 7 可以看出,在不断增加规则特征的情况下,准确率和召回率虽有些上下波动,但 *F* 值在不断地提升,说明本文提出的规则特征可以提高指代

消解实验的性能。

3.2.5 IndRNN 层数对实验结果的影响

在本实验中用以提取全局特征的 IndRNN 采用的是两层结构,理论上堆叠的层数越多,就可以得到更深层次、更加抽象的语义信息。因此本文验证 IndRNN 层数对实验结果的影响,实验结果如表 8 所示。

表 8 IndRNN 层数对结果的影响

层次	<i>P</i> /%	<i>R</i> /%	<i>F</i> /%
1	83.61	86.3	84.95
2	87.23	88.8	88.04
3	85.18	89.67	87.37
4	85.55	87.25	86.39

由表 8 可以看出,随着层数的增加,准确率、召回率和 *F* 值都是先增加再减小,并且在层数为 2 的时候准确率和 *F* 值达到最大,而召回率在层数为 3 时达到最大。所以本文将 IndRNN 的层数设置为 2,以便取得更好的效果。

4 总结

本文提出一种基于注意力机制的混合模型的维吾尔语名词指代消解方法,通过引入注意力机制将特征内在的权重计算出来,进而分别利用 IndRNN 和 CNN 得到富含上下文信息的全局特征和局部特征;再将两种特征进行融合,进而可以得到更好的结果。另外在规则特征的前提下引入了语义特征,可以得到先行语和照应语在实验文本中的深层次语义信息,进一步提高特征的代表性。实验结果证明,该方法对于维吾尔语指代消解有较好的效果,可以明显提高实验的性能,并且在引入语义特征后可以显著提高实验的效果。

参考文献

[1] Van deemter K. Kibble R. On coreferring: Coreference in MUC and related annotation schemes [J]. Computational Linguistics,2006,26(4): 629-637.

[2] 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.

[3] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报, 2002, 16(6):9-17.

- [4] Mc Carthy J F, Lehnert W G. Using decision trees for coreference resolution[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, 47(1):1050-1055.
- [5] Soon W M, Ng H T, Lim D C Y. A machine learning approach to coreference resolution of noun phrases [M]. MIT Press, 2001.
- [6] Ng V, Cardie C. Improving machine learning approaches to coreference resolution[C]//Processing of the 40th Annual Meeting on Association for Computational Linguistics. 2002:104-111.
- [7] Yang X, Su J, Tan C L. A Twin-candidate model for learning-based anaphora resolution[J]. Computational Linguistics, 2008, 34(3):327-356.
- [8] Kong F, Zhou G D, Zhu Q. Employing the centering theory in pronoun resolution from the semantic perspective[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2. 2009:987-996.
- [9] 郭志立. 人称代词指代主体的辨析及其在摘要提取中的应用[C]. 中文信息处理国际会议, 1998.
- [10] 马彦华, 王能忠. 汉语中人称代词指代问题研究 [C]. 中文信息处理国际会议, 1998.
- [11] 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(6): 633-637.
- [12] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报, 2002, 16(6):9-17.
- [13] 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.
- [14] 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.
- [15] 李冬白, 田生伟, 禹龙, 等. 基于深度学习的维吾尔语人称代词指代消解[J]. 中文信息学报, 2017, 31(4):80-88.
- [16] 李敏, 禹龙, 田生伟, 等. 基于深度学习的维吾尔语名词短语指代消解[J]. 自动化学报, 2017(11): 1984-1992.
- [17] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.
- [18] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention [C]//Proceedings of NIPS 2014, 2014, 3:2204-2212.
- [19] Bandanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv: 2014.
- [20] Li S, Li W, Cook C, et al. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN[J]. arXiv preprint arXiv:1803.04831, 2018.



祁青山(1994—), 硕士研究生, 主要研究领域为自然语言处理。
E-mail: QQS_XJ_U@163.com



禹龙(1974—), 硕士, 教授, 主要研究领域为计算机智能技术和计算机网络。
E-mail: yul_xju@163.com



田生伟(1973—), 通信作者, 博士, 教授, 主要研究领域为计算机智能技术和自然语言处理。
E-mail: tianshengwei@163.com