基于 Bi-LSTM 的维吾尔语人称代词指代消解

田生伟[1],秦 越[2],禹 龙[3],吐尔根·依布拉音[2],冯冠军[4]
(1. 新疆大学软件学院,新疆乌鲁木齐 830008; 2. 新疆大学信息科学与工程学院,新疆乌鲁木齐 830046;
3. 新疆大学网络中心,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: 针对维吾尔语人称代词指代现象,提出利用双向长短时记忆网络(Bi-directional long short term memory, Bi-LSTM)的深度学习机制进行基于深层语义信息的维吾尔语人称代词指代消解.首先将富含语义和句法信息的 word embedding 向量作为 Bi-LSTM 的输入,挖掘维吾尔语隐含的上下文语义层面特征;其次对维吾尔语人称代词指代现象进行探索,提取针对人称代词指代研究的 24 个 hand-crafted 特征;然后利用多层感知器(multilayer perception, MLP)融合 Bi-LSTM 学习到的上下文语义层面特征与 hand-crafted 特征;最后使用融合的两类特征训练 softmax 分类器完成维吾尔语人称代词指代消解任务.实验结果表明,充分利用两类特征的优势,维吾尔语人称代词指代消解的 F1 值达到 76.86%.实验验证了 Bi-LSTM 与单向 LSTM、浅层机器学习算法的 SVM 和 ANN 相比更具备挖掘隐含上下文深层语义信息的能力,而 hand-crafted 层面特征的引入,则有效提高指代消解性能.

关键词: 指代消解; 双向长短时记忆网络; 词向量; 深度学习; 维吾尔语; 自然语言处理
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2018)07-1691-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.07.022

Anaphora Resolution of Uyghur Personal Pronouns Based on Bi-LSTM

TIAN Sheng-wei[1], QIN Yue[2], YU Long[3], Turgun Ibrahim[2], FENG Guan-jun[4]
(1. *College of Software, Xinjiang University, Urumqi, Xinjiang* 830008, *China*;
2. *College of Information Science and Technology, Xinjiang University, Urumqi, Xinjiang* 830046; *China*;
3. *Network Center, Xinjiang University, Urumqi, Xinjiang* 830046, *China*;
4. *College of Humanities, Xinjiang University, Urumqi, Xinjiang* 830046, *China*)

Abstract: Specific to the anaphora phenomenon of Uyghur personal pronouns, a deep learning mechanism of Bi-LSTM (Bi-directional long short term memory) network is proposed, which is based on the deep semantic information to resolve anaphora resolution problem in Uyghur personal pronouns. Firstly, make the word embedding which contain semantic and syntactic information as the input of Bi-LSTM, to excavate the implicit semantic features of Uyghur. Secondly, explore the anaphora phenomenon in Uyghur and extract 24 hand-crafted features. Then, use multilayer perception(MLP) to concatenate hand-crafted features and context semantic features. Finally, two types of features are used to train the softmax classifier to complete the task. The experimental results show that, on the basis of full utilization of the advantages of two types of features, the F1 value of anaphora resolution is 76.86%. It is proved that Bi-LSTM is more capable of mining implicit context semantic information than LSTM、SVM as well as ANN, and the introduction of hand-crafted features can effectively improve the performance.

Key words: anaphora resolution; Bi-LSTM; word embedding; deep learning; Uyghur; natural language processing

1 引言

指代(anaphroa)是指在语篇中用一个指代词回指某个前文出现的语言单位,对简化表述、衔接上下文、保持语篇连贯性起重要作用,将指代成分无歧义的消解有利于机器分析及篇章理解[1].在语言学中,用于指向的指代词称为照应语(anaphora),所指的内容或对象称为先行语(antecedent),指代消解就是确定照应语和先

收稿日期:2017-03-30;修回日期:2017-09-09;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61563051, No.61662074, No.61262064);国家自然科学基金重点项目(No.61331011);新疆维吾尔自治区科技人才培养项目(No.QN2016YX0051)

行语相互联系的过程[2],而维吾尔语人称代词指代消解主要考察人称代词与邻近名词性短语的指代关系.例如(维吾尔语的书写习惯为从右向左):

ئىسمايىل مېھمان سارىيىغا قايتىپ كەتتى،

ئۇ ئىشىكلىدە قوغدىغۇچىلىرى بارئىسىل ،ھەشەمەتلىك مېھمانخاندا يالغۇز ياتاتتى.

ياتاق ئىشىكىنى كىمدە كىم چەكمىسۇن ،ئۇ زادىلا ئاچمايتتى

(阿不都热合曼·卡哈尔:《如诉的歌》383 页)

(译:伊斯玛依江回到旅馆,他独自住在一个门口有护卫并且很豪华的旅馆.无论谁敲门,他也绝对不会开.)

由例句可知,ئۇ(他)作为照应语指代前文中的ئىسمايىل(伊斯玛依江),ئۇ(他)与ئىسمايىل(伊斯玛依江)存在人称代词与名词短语的指代关系,而与ئۇ(他)相邻的名词短语还有:ئىشىكىنى(门)、سارىيىغا(旅馆)、قوغدىغۇچىلىرى(护卫)、ئىشىكلىدە(门口).如何让机器正确的判断维吾尔语人称代词与哪个相邻名词短语之间存在指代关系,是本文探索的主要问题.

近年来,基于机器学习的方法在指代消解研究中得到广泛运用.在有关英文指代消解研究中,文献[3]将判断先行语的问题转换为二分类问题,通过分类器确定照应语和先行语之间是否具有指代关系;文献[4]依据这一思想,提出一个完整的基于机器学习的指代消解平台;文献[5]在文献[4]研究的基础上,对词法、语法及语义方面进行探索,使指代消解性能更优;文献[6]通过基于机器学习的英文名词短语指代消解平台探索了距离特征对指代消解性能的影响.在有关中文指代消解研究中,文献[7]采用卷积树核函数,从多方面对结构化句法树进行动态扩展,从而提升代词消解的性能;文献[8]提出一种最优测度 Laplacian SVM 算法解决了中文指代消解语料不足的问题;文献[9]提出利用深度学习机制进行基于深层语义的代词指代消解,研究深层语义信息在指代消解中的作用.由前人工作可知,基于机器学习的指代消解研究有效提高了指代消解性能,而对深度学习机制的探索,则使模型能够学习出文本中潜在的多层表达和深层的语义信息.

目前,指代消解研究对上下文深层语义方面探索较少,且主要集中在中文和英文等大语种,对维吾尔语等少数民族语言的研究不够深入.针对此问题,本文探索具有典型指代特色的维吾尔语人称代词消解问题,采用 word embedding[10]表示词序序列中照应语和候选先行语及其上下文词汇的特征向量,将其作为双向长短时记忆网络(Bi-directional long short term memory,Bi-LSTM)的输入,利用 Bi-LSTM 捕获每个词汇前向及后向上下文信息的优势,挖掘照应语与候选先行语之间隐含的深层语义关系,由此在上下文语义层面揭示照应语与候选先行语之间的联系.此外,根据维吾尔语人称代词指代特点,提取针对人称代词指代研究的 hand-crafted 层面特征,与 Bi-LSTM 学习到的上下文语义层面特征融合.一方面利用基于 Bi-LSTM 的深度学习机制自动学习潜在的句法和语义特征,另一方面借助规则表示的人工提取特征刻画维吾尔语人称代词指代现象.得益于上述两类特征的优势,最终有效完成维吾尔语人称代词指代消解任务.

2 相关研究

鉴于深度学习在图像和计算机视觉等领域取得重大突破,学者们尝试在自然语言处理领域运用深度学习技术.其中,序列化模型将文本视为有序词汇序列,充分考虑文本的有序性和词汇间的关联性,更加符合自然语言规律.在解决序列化问题时,循环神经网络(Recurrent neural network,RNN)可充分利用上下文信息,然而,RNN 在训练过程中存在梯度爆炸和消失的问题[11],由文献[12]提出的长短时记忆网络(long short term memory,LSTM)则有效地解决了这一问题.在有关研究中,文献[13]借助 Bi-LSTM 从 word embedding 中获取到高层面特征,以完成关系分类问题;文献[14]利用 Bi-LSTM 能够抓住目标词及其上下文联系解决情感分类问题.

对于文本特征表示方面,Bengio 等[15]使用 word embedding 将词汇向量化表示,替代传统 one-hot 表达,解决了 one-hot 带来维数灾难的问题.此外,学者们还发现,通过大规模无标记语料库训练出的 word embedding 蕴含丰富的语义和句法信息,且能够表达词汇间的相似度.文献[16]将 word embedding 用于词性标注、命名实体识别、语义角色标注等任务,以无监督的方式训练模型并将 word embedding 向量作为初始值,从而促进后续有监督模型的性能提高.这种将 word embedding 作为模型输入的使用方式被后续自然语言处理任务广泛采用.

3 问题形式化描述

3.1 指代关系

设 RealWorld 为"真实世界"的对象集合,对象 $r \in$ RealWorld,若自然语言表达式 x 指代某一对象 r,则表示为 denote(x) = r,即 x 与 RealWorld 集合中的 r 存在指代关系.例如在引言的例句中,x = ئۇ(他),r = ئىسمايىل(伊斯玛依江),denote(ئۇ(他)) = ئىسمايىل(伊斯玛依江),则认为ئۇ(他)与ئىسمايىل(伊斯玛依江)存在指代关系.

3.2 可被人称代词消解的实体

本文将能够被人称代词所指代的实体集合表示为

{Obj-Pronoun},{Obj-Pronoun} = 人名集合∪地名集合∪组织机构名集合∪表示人的普通名词集合∪人称代词集合.

例如,维吾尔语人称代词作为照应语,需要消解的实体为集合{Obj-Pronoun}中的对象,即候选先行语. 通过对维吾尔语人称代词指代现象的研究,实验组维吾尔语语言学专家总结出可能的先行语集合为:人名集合{ئىسمايىل(伊斯玛依江),ساۋۈت(沙吾提),…};组织机构名集合{تەشكىلاتى دۇنيا ساھىيە(世界卫生组织),…};地名集合{ئۈرۈمچى(乌鲁木齐),…};表示人的普通名词集合{ئوقۇغۇچىلار(学生们),ئامېرىكا رەھبىرى(美国领导),…}.

3.3　指代链

设在语篇中有 $NP_1, NP_2, NP_3, \cdots, NP_n$ 的名词短语词序序列,其中,若 NP_i 与 NP_j 存在指代关系,即 denote(NP_i) = NP_j,则称 NP_i 与 NP_j 在同一条指代链中,表述为 IsChain(NP_i, NP_j) = True;反之,若 NP_i 与 NP_j 不存在指代关系,即 NP_i 与 NP_j 不在同一条指代链中,表述为 IsChain(NP_i, NP_j) = False.

例如在引言的例句中,ئىسمايىل(伊斯玛依江)与ئۇ(他)存在指代关系,即在同一条指代链中,表述为 IsChain(ئىسمايىل(伊斯玛依江),ئۇ(他)) = True.

3.4　维吾尔语人称代词语法特点

维吾尔语为一种黏着型语言,属阿尔泰语系突厥族语葛逻禄语支,是我国语言的重要组成部分之一. 维吾尔语形态变化多样,语法形式复杂,其通过在词语的词尾缀接不同的词缀来表示不同的语法功能. 例如:تۇرسۇنىچە(吐尔逊觉得),词缀"نىڭچە"缀接到人名"تۇرسۇن(吐尔逊)"后,表达了"吐尔逊觉得"的意思.

维吾尔语人称代词分为第一人称代词,例如مەن(我);第二人称代词,例如سەن(你);第三人称代词,例如ئۇ(他、她、它). 维吾尔语人称代词与汉语(或英语)最大的区别在于维吾尔语第三人称代词没有性别之分,如英语第三人称单数有"he/she/it"之分,而在维吾尔语中,第三人称既可指男性,也可指女性,还可指物体,因此第三人称较第一人称和第二人称相比应用场景更广泛,指代现象更频繁.

维吾尔语人称代词的特性主要受"格"的形式影响. "格"语法[17]是一种特殊的语言形式,"格"形式不同,附加的"格后缀"也不同. "格"语法体现出名词性短语在篇章语句中的句法功能,在语法形式上具备独立性,语法意义上具备稳定性,它作为维吾尔语人称代词的重要语言特征之一,为人称代词指代消解研究提供依据. "格"语法包括主格、属格、向位格等十种形式.

4　基于 Bi-LSTM 的人称代词指代消解

本文设计了维吾尔语人称代词指代消解框架,其核心思想是采用双层的 Bi-LSTM 挖掘照应语和候选先行语的上下文语义特征,此外探索基于维吾尔语人称代词指代现象的 24 个 hand-crafted 特征. 使用 MLP 将上述两类特征融合,由 softmax 进行分类判断照应语与候选先行语之间的关系. 图 1 为维吾尔语人称代词指代消解框架及模型结构,其中,上部分为指代消解框架,下部分为框架中具体的模型结构部分.

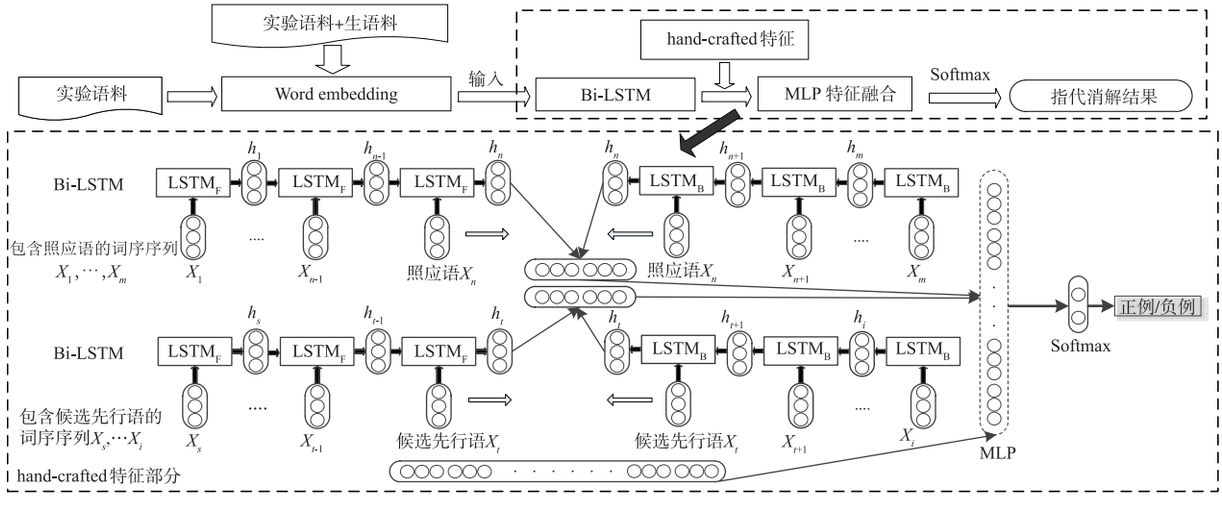

图1　基于Bi-LSTM的维吾尔语人称代词指代消解框架及模型结构

4.1　生成训练实例和测试实例

在本文中,首先使人称代词与它之前出现过的名词短语依次进行配对,生成训练/测试实例.

生成训练实例时,指代链信息已知. 对识别出的维吾尔语人称代词,查找其是否在某条指代链中,若不在,则视为非待消解项,不必为其寻找合适的先行语;若在某条指代链中,则将该人称代词视为照应语 X_n,为

其寻找合适的先行语. 经维吾尔语语料统计,选取与照应语 X_n 距离为四句之内的名词短语 NP_0, NP_1, …, NP_n,依次进行配对. 若存在 $NP_{i(0<i<n)}$,使 IsChain(X_n, NP_i) = True,则名词短语对〈X_n, NP_i〉为正例;而照应语 X_n 与 NP_{i+1},, NP_n 配对的名词短语对〈X_n, NP_{i+1}〉…〈X_n, NP_n〉为负例.

生成测试实例时,指代链信息未知. 对识别出的人称代词均视为照应语,选取与它四句之内的名词短语,配对为〈照应语,候选先行语〉对,交由模型判断,是否存在指代关系.

4.2 维吾尔语人称代词指代消解模型结构

4.2.1 LSTM

LSTM 能够对时间序列进行显式建模,这一特点更加适用于语句中词序序列的语义表达,且符合自然语言的表现形式. 如图 2 所示,LSTM 由记忆单元组成,记忆单元由 3 个控制门控制 LSTM 的记忆信息,这 3 个门分别为:输入门(input gate)t_i,遗忘门(forget gate)f_t,输出门(output gate)o_t.

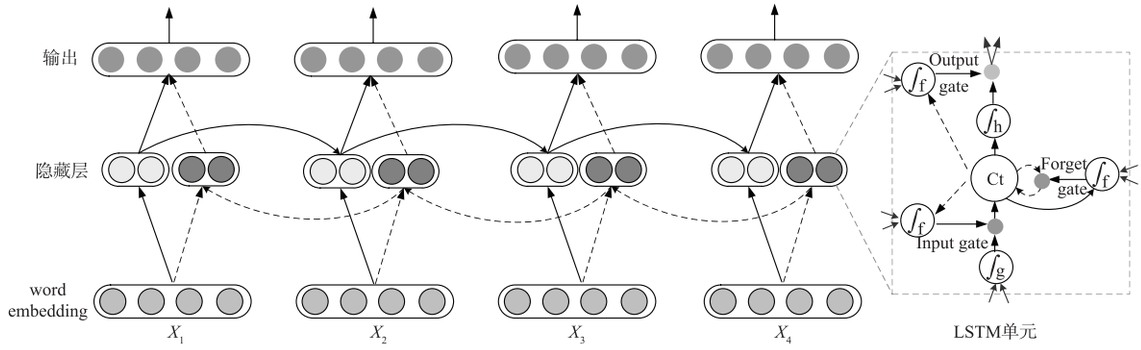

图2 LSTM网络

记忆单元能够记忆某一时间段的信息,对于指代消解问题,利用 LSTM 记忆单元来记忆语句中之前某一时刻的词汇信息,充分学习篇章上下文词汇序列,从而更深入的发掘语句词汇序列中的语义及句法特征.

设输入的词序序列为 $x = x_1, x_2, \cdots, x_n$ 的 word embedding 向量,在时刻 t,LSTM 单元对应输入词汇 X_t 和前一个 LSTM 单元的输出 h_{t-1},则在 t 时刻 LSTM 单元表述为:

输入门:$i_t = \sigma(W_{x_i}x_t + W_{h_i}h_{t-1} + b_i)$ (1)

遗忘门:$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$ (2)

输出门:$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$ (3)

记忆单元:$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$ (4)

单元输出:$h_t = o_t \tanh c_t$ (5)

在上式中,σ 为激活函数 sigmoid;参数 $W*$, $b*$ 分别为模型权值和偏置系数,i,f,o 分别代表输入门,遗忘门,输出门和记忆单元.

从 LSTM 模型结构角度来看,其单向机制仅包含照应语(候选先行语)的前文信息,而对后文信息并未涉及. 这一特性是否影响指代消解性能,将在实验部分的 5.3.1 节进行验证.

4.2.2 Bi-LSTM

维吾尔语人称代词指代消解模型结构由两层 Bi-LSTM 网络组成,其输入均为语句中的词序序列. 如图 1 所示,一层为包含候选先行语的 Bi-LSTM,其输入为目标候选先行语以及其相邻前 $t-s$ 后 $i-t$ 个词汇的 word embedding 向量;另一层为包含人称代词(照应语)的 Bi-LSTM,相同地,其输入为目标照应语及其相邻前 n 个和后 $m-n$ 个词汇的 word embedding 向量.

Bi-LSTM 是常见的考虑上下文信息的模型之一,其基本思想是分别使用单独的前向 LSTM($LSTM_{Forward}$,下文简称 $LSTM_F$)及后向 LSTM($LSTM_{Backward}$,下文简称 $LSTM_B$),进行前向传播和反向传播两个过程. $LSTM_F$ 的输入为照应语(候选先行语)之前词汇的 word embedding 向量,$LSTM_B$ 的输入为照应语(候选先行语)之后词汇的 word embedding 向量,由语句的词序序列从前至后运行 $LSTM_F$,再从后至前运行 $LSTM_B$,从而挖掘照应语(候选先行语)前文和后文的隐含语义,以表示照应语(候选先行语)的上下文信息. 两个单独的 $LSTM_B$ 和 $LSTM_F$ 连接在共同的输出层,意味着对于语句中的词序序列 $x = x_1, x_2, \cdots, x_n$,整个 Bi-LSTM 包含照应语(候选先行语)完整且连续的前文和后文语境信息.

由上文可知,Bi-LSTM 包含 $LSTM_F$ 和 $LSTM_B$ 两个子网络,则将 Bi-LSTM 的输出表示为:

$$Fi = LSTM_{F\_h} \oplus LSTM_{B\_h} \quad (6)$$

其中,$LSTM_{F_h}$代表前向的 $LSTM_F$ 最后一个单元的隐藏状态输出,$LSTM_{B_h}$代表后向的 $LSTM_B$ 最后一个单元的隐藏状态输出.

4.3 维吾尔语人称代词 hand-crafted 特征

基于维吾尔语指代现象的 hand-crafted 层面特征,凸显了人称代词指代方面的知识表示,刻画了维吾尔语人称代词指代现象. 本文提出的 hand-crafted 特征依据国内外指代消解研究及维吾尔语语言特性,从下述五个角度进行筛选,主要包括(在特征表述中,设 X_n 为照应语,X_t 为候选先行语):

(1)词汇特征方面

①照应语为代词(AnaphorPron.):该特征表示为 $v_{ap}=\{0,1\}$. 若 X_n 为代词,特征值取 1,否则取 0.

②先行语为代词(CandidatePron.):该特征表示为 $v_{cp}=\{0,1\}$. 若 X_t 为代词,特征值取 1,否则取 0.

③是否存在嵌套(NestPron.):该特征表示为 $v_{NestPron.}=\{NestPron(X_n),NestPron(X_t)\}$. NestPron($X_n$)表示 X_n 是否嵌套在其他名词短语内,若是取 1,否则取 0;NestPron(X_t)表示 X_t 是否嵌套在其他名词短语内,若是取 1,否则取 0.

(2)语义和句法特征方面

①语义角色(SemanticRole.):该特征表示为 $v_{role}=\{arg_0(X_n),arg_1(X_n),arg_0(X_t),arg_1(X_t)\}$. $arg_0(j)$ 表示 j 的语义角色是否为施事者,若是取 1,否则取 0;$arg_1(j)$ 表示 j 的语义角色是否为受事者,若是取 1,否则取 0.

②语义类别一致性(SemanticAgree.):该特征表示为 $v_{sa}=\{0,0.5,1\}$. 若 X_n 和 X_t 语义类别一致,特征值取 1;若不一致,则取 0;若 X_n 和 X_t 其中一个为未知,取 0.5.

③语义关系(SemanticRelation.):该特征表示为 $v_{SemRela.}=\{Agt(Xn),Pat(Xn),Agt(Xt),Pat(Xt)\}$. Agt($j$)表示 j 的语义关系是否存在施事关系,若存在取 1,否则取 0;Pat(j)表示 j 的语义关系是否存在受事关系,若存在取 1,否则取 0.

④句法关系(SyntaticRelation.):该特征表示为 $v_{SynRela.}=\{SBV(Xn),VOB(Xn),SBV(Xt),VOB(Xt)\}$. $SBV(j)$ 表示 j 的句法关系是否存在主谓关系,若存在取 1,否则取 0;VOB(j)表示 j 的句法关系是否存在动宾关系,若存在取 1,否则取 0.

(3)语法特征方面

①生命度一致性(AnimateAgree.):该特征表示为 $v_{aa}=\{0,1\}$. 若 X_n 和 X_t 生命度一致,特征值取 1,否则取 0. 在维吾尔语人称代词指代中,候选先行语不仅是有生命的人、动物等,还可以是无生命的实体、机构等.

②性别一致性(GenderAgree.):该特征表示为 $v_{ga}=\{0,0.5,1\}$. 若 X_n 和 X_t 性别一致,特征值取 1;若不一致,则取 0;若 X_n 和 X_t 其中一个为未知,则取 0.5.

③单复数一致性(NumberAgree.):该特征表示为 $v_{na}=\{0,0.5,1\}$. 若 X_n 和 X_t 单复数一致,特征值取 1;若不一致,则取 0;若 X_n 和 X_t 其中一个为未知,取 0.5.

(4)位置及距离特征方面

①位置匹配一致性(PositionAgree.):该特征表示为 $v_{PositionAgree.}=\{0,0.5,1\}$. 位置匹配主要考察 X_n 和 X_t 是否位于语句的开头. 在人称代词指代现象中,所指代的人、物或实体,有倾向指代语句的主语. 维吾尔语中正常语序是“主语 + 宾语 + 谓语”,即主语在语句的开头. 若 X_n 和 X_t 均在语句的开头,特征值取 1;若只有其中一个在语句的开头,特征值取 0.5,若均未在语句的开头,则取 0.

②距离特征(Distance):该特征表示为 $v_{Distance}$. 距离特征是 X_n 和 X_t 语句的空间距离. 空间距离越近,则 X_n 和 X_t 存在指代关系的可能性越大. 对空间距离进行逆向取值,并归一化在 0 到 1 之间. 定义 $v_{Distance}(d)$ 为该特征取值,设空间距离为 d,若 d 大于等于 10 句,则特征值取 0;若 d 小于 10 句,则特征值取 $0.1\times(10-d)$,则

$$v_{Distance}(d)=\begin{cases}0, & d\geqslant 10\\ 0.1\times(10-d), & 0\leqslant d<10\end{cases} \tag{7}$$

(5)维吾尔语特性方面

①可被人称代词消解的集合(PronResolution):该特征表示为 $v_{PronResolution}=\{0,1\}$. 若 X_t 出现在 3.2 节的{Obj-Pronoun}集合中,则特征值取 1,否则取 0.

②“格”语法一致性(CaseAgree.):该特征表示为 $v_{CaseAgree.}=\{0,0.5,1,2\}$. 若 X_n 和 X_t“格”语法一致,特征值取 1;若不同,特征值取 0;若其中一个为无“格”语法,特征值取 0.5;若 X_n 和 X_t 均为无“格”语法,则取 2.

根据上述五个角度的对维吾尔语人称代词指代现象的探索,hand-crafted 层面特征如表 1 所示,以在引言中的例句中照应语ئۇ(他)与候选先行语ئىسمايىل(伊斯玛依江)为例.

表 1　维吾尔语人称代词 hand-crafted 层面特征样例

照应语	先行语	hand-crafted 层面特征													
		词汇特征			语义和句法特征				语法特征			位置及距离特征		维语特性	
ئۇ	ئىسمايىل	V_{ap}	V_{cp}	V_{NestP}	V_{Role}	V_{sa}	$V_{SemRela}$	$V_{SynRela}$	V_{aa}	V_{ga}	V_{na}	$V_{PosAgr.}$	V_{Dis}	$V_{PronRe.}$	$V_{CaseAgr}$
(他)	(伊斯玛依江)	1	0	0,0	0,0,0,0	1	0,0,0,0	1,0,1,0	1	1	1	1	0.7	1	1

4.4 特征融合与 softmax 分类

利用多层感知器(multilayer perception,MLP)将 Bi-LSTM 学习到的照应语和候选先行语上下文隐含深层语义特征,与针对维吾尔语人称代词指代研究的 hand-crafted 特征进行融合,接着使用 softmax 分类器进行分类,从而确定照应语和候选先行语之间的关系.

在本文中,若照应语 X_n 与候选先行语 X_t 存在 IsChain(X_n,X_t) = True 的表述,则判断 X_n 与 X_t 关系时为正例,标签为 1;若照应语 X_n 与先行语 X_t 存在 IsChain(X_n,X_t) = False 的表述,则判断 X_n 与 X_t 关系时为负例,标签为 0.

5 实验与分析

5.1 语料来源

目前为止,国际上通用的指代消解标注语料以 MUC(仅有 English 语料)、ACE(有 Arabic、English、Chinese 三种语料)为主,尚未发现公开通用的维吾尔语指代消解评测语料库,因此,本文对语料进行筛选和标注以完成维吾尔语人称代词指代消解研究.

实验语料以天山网、人民网、昆仑网等维吾尔语版网页为来源,使用网络爬虫或人工下载页面,对页面内容进行去重处理,最终获取实验所需的原始文本. 为确保语料的普遍性,选取的语料内容包含新闻、小说等. 在实验组维吾尔语语言学专家的指导下,对语料进行人工标注并选用 XML 文件存储. 所选标注完成指代链等信息的语料共 179 篇,其中包含指代关系的〈名词短语,人称代词〉1421 对,不包含指代关系的〈名词短语,人称代词〉3615 对,共 5036 对. 表 2 为其中人称代词类型的分布情况.

表 2 维吾尔语料中人称代词类型分布

人称代词类型	所占数目	所占比例
第一人称	270	32.42%
第二人称	87	10.44%
第三人称	476	57.14%
总 计	833	100%

5.2 实验评测标准

本文实验测评方式采用自然语言处理研究中常用的 MUC 标准,即准确率 P,召回率 R 和 $F1$ 值考察指代消解性能. 其中,P 指正确消解的实体占实际消解实体的百分比;R 指正确消解的实体占指代消解系统应消解实体的百分比;$F1$ 值为准确率和召回率的综合评价指标,即:$F1 = 2 \times R \times P/(R + P)$.

5.3 实验设计

为探索不同角度下维吾尔语人称代词指代消解的性能,本文设计了如下 5 个实验:(1)Bi-LSTM 与其他模型的指代消解性能对比;(2)word embedding 维度对指代消解性能的影响;(3)验证 hand-crafted 层面特征对指代消解性能的影响;(4)语料规模对指代消解性能的影响;(5)人称代词子类指代消解性能对比.

此外,本文对原有实验语料进行扩充,从大型维吾尔语网站获取共约 7000 余篇题材丰富的生语料,进行去重、去噪处理后,使用 Mikolov[18] 提出的 word2vec 工具,选择 CBOW 模型作为训练框架,训练 k 维(k = 50,100,150,200,250)的 word embedding 向量.

在实验过程中,将文本的词序序列先选用 200 维的 word embedding 向量作为模型的输入数据. 为避免实验不确定性,确保数据随机性,本文实验均采取五折交叉验证法进行,经过反复尝试网络模型的不同参数组合,确定了基于本实验数据量下的最优参数,如表 3 所示.

表 3 模型最优参数

参数	值
学习率	0.01
迭代次数	15
LSTM 隐藏层节点数目	110
MLP 隐藏层节点数目	100
MLP 层数目	4
参数更新算法	adagrad

5.3.1 Bi-LSTM 与其他模型的指代消解性能对比

将第 4 节图 1 框架(下文称为基准框架)中的 Bi-LSTM 模型换为 LSTM 模型,并与基准框架进行指代消解性能对比. 此外,Bi-LSTM 还与浅层机器学习的支持向量机(Support Vector Machine,SVM)、人工神经网络(Artificial Neural Network,ANN)进行指代消解性能对比. 其中,SVM 使用 RBF 核函数,γ = 1.5;ANN 为三层网络结构,隐藏层节点数为[120,100,95],迭代次数为 150,批尺寸为 90,实验结果如表 4 所示.

表 4 Bi-LSTM 与其他模型的指代消解性能对比

模型	P/%	R/%	$F1$/%
Bi-LSTM	82.33	72.07	76.86
LSTM	77.08	69.81	73.27
SVM	67.15	74.19	70.49
ANN	70.93	71.77	71.34

由表 4 可知,浅层机器学习的 SVM 和 ANN 与 Bi-LSTM 和 LSTM 相比,反映整体性能的 $F1$ 值均有所下降,这是由于 SVM 和 ANN 挖掘数据中隐藏深层语义信息的能力与 Bi-LSTM 和 LSTM 相比相对较弱,Bi-LSTM 和 LSTM 的深度学习机制能够捕捉数据中的复杂分布,学习出文本中隐含的高层特征. 另外,Bi-LSTM 与 LSTM 相比,反映整体性能的 $F1$ 值提升了 3.59%. 在本研究中,Bi-LSTM 的性能优于 LSTM,这是因为从模型角度看,Bi-LSTM 双向机制的模型结构可挖掘照应语和候选先行语丰富的前文和后文语义信息,而单向的 LSTM 只

能捕捉到照应语和候选先行语的前文信息,对后文信息并未涉及. 由上述分析可知,维吾尔语人称代词指代消解研究使用双向机制下的 Bi-LSTM 可挖掘照应语(候选先行语)隐含的上下文深层语义信息,Bi-LSTM 与 LSTM、SVM、ANN 相比更适用于本任务,因此在本研究数据集下的后续实验选择 Bi-LSTM 作为学习上下文语义层面特征的模型.

5.3.2　word embedding 维度对指代消解性能的影响

word embedding 维度是生成 word embedding 向量的参数之一,为了考察不同维度对指代消解性能的影响,训练维度为 50 维、100 维、150 维、200 维、250 维的 word embedding 向量作为基准框架中 Bi-LSTM 的输入,实验结果如表 5 所示.

表 5　不同维度下指代消解性能对比

维度	*P*/%	*R*/%	*F*1/%
50	69.17	59.54	64.00
100	70.10	69.20	69.65
150	75.59	68.08	71.64
200	82.33	72.07	76.86
250	83.11	71.11	76.64

由表 5 可知,随着 word embedding 维度的增加,准确率 *P* 逐渐增大,在 250 维达到最优;反映整体性能的 *F*1 值也随着维度的增加而增大,在 200 维达到最优,而 250 维与 200 维相比,*F*1 值降低了 0.18%,并未显著增加. 在 200 维时 *F*1 值最优,这是由于高维度向量蕴含比低维度向量更丰富的语义信息,使模型能够有效的学习到数据内部分布的语义表达;而在 250 维的 *F*1 值并未显著增加,这是由于高维度向量中包含更多的信息,但其中也包含一些无用的干扰信息和噪音,影响指代消解的性能. 因此在本研究数据集下的后续实验选择 200 维 word embedding 向量.

5.3.3　hand-crafted 特征对指代消解性能的影响

以上实验均结合了 Bi-LSTM 学习到的上下文语义特征和依据维吾尔语指代现象的 hand-crafted 特征. 为探索 hand-crafted 特征对指代消解性能的影响,将去掉 hand-crafted 特征的模型与包含全部两类特征的模型作对比. 此外,本节还探索了 hand-crafted 特征中的维吾尔语特性特征($V_{\text{PronResolution}}$ 和 $V_{\text{CaseAgree.}}$)对指代消解性能的影响,使包含 hand-crafted(去 $V_{\text{PronResolution}}$ 和 $V_{\text{CaseAgree.}}$)特征的基准框架与包含全部 hand-crafted 特征的基准框架做对比,实验结果如表 6 所示.

表 6　hand-crafted 层面特征对指代消解性能的影响

特征类型	*P*/%	*R*/%	*F*1/%
$F_{\text{hand-crafted}}+F_{\text{上下文语义}}$	82.33	72.07	76.86
$F_{\text{hand-crafted(去VPronResolution和VCaseAgree.)}}+F_{\text{上下文语义}}$	81.93	70.18	75.61
$F_{\text{仅包含上下文语义}}$	74.44	63.27	68.40

由表 6 可知,在去掉 hand-crafted 特征,仅包含上下文语义特征条件下,其准确率 *P* 与包含全部两类特征的准确率 *P* 相比降低了 7.89%,反映整体性能的 *F*1 值降低了 8.46%. 本文选择引入 hand-crafted 特征,因为 hand-crafted 特征针对维吾尔语人称代词指代现象,是根据维吾尔语语言特色总结并提取的特征集,生动反映了维吾尔语中照应语和候选先行语在知识和规则方面的联系与表示. 实验结果表明,去掉 hand-crafted 特征后,指代消解性能下降,证实了 hand-crafted 特征的引入,对本研究任务的有效性. 而去掉 $V_{\text{PronResolution}}$ 和 $V_{\text{CaseAgree.}}$ 特征的基准框架与包含全部特征的基准框架相比,其 *F*1 值下降了 1.25%,说明 hand-crafted 特征中的维吾尔语特性特征($V_{\text{PronResolution}}$ 和 $V_{\text{CaseAgree.}}$)对指代消解性能有影响. $V_{\text{PronResolution}}$ 和 $V_{\text{CaseAgree.}}$ 作为最具维吾尔语语言特色的特征,将二者引入,确定了候选先行语的范围,并且在“格”语法方面体现了照应语和候选先行语间的联系.

5.3.4　语料规模对指代消解性能的影响

语料规模对实验性能也有影响,若语料规模太小,则包含的数据量较少,模型能够挖掘数据中隐含的语义信息也相应较少,这可能会对指代消解性能造成影响. 针对此问题,本节在逐步扩大语料规模的情况下进行实验. 为方便描述语料规模,采用〈名词短语,人称代词〉短语对的数目作为语料规模大小的表述,随机抽取包含 *t* 对的〈名词短语,人称代词〉短语对,并逐步扩大短语对规模,使用包含上述短语对的语序序列进行 20 组实验,实验结果如图 3 所示.

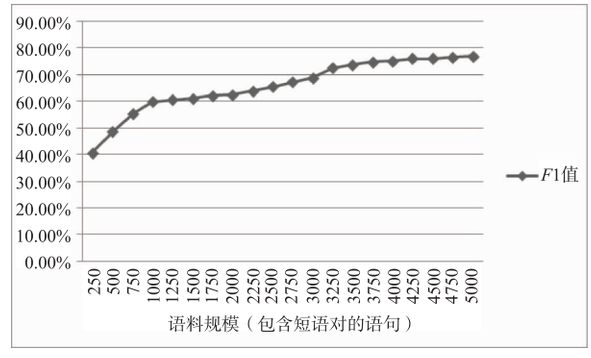

图3　语料规模对指代消解性能影响

由图 3 可知,反映整体性能的 *F*1 值随着语料规模的增大,在 250 ~ 1000 对的区间内迅速上升,在 1000 ~ 3750 对的区间内上升缓慢,在 3750 ~ 5000 对的区间内趋势逐渐平稳. *F*1 值在一定语料规模时趋于稳定,不再波动,说明本文所选语料适中,满足此实验需求.

5.3.5　人称代词子类指代消解性能

第 3.4 节已介绍维吾尔语人称代词分类及特点,为考察人称代词子类指代消解性能差异,分别探究了第

一人称、第二人称、第三人称的指代消解性能,实验结果如表7所示.

表7 人称代词子类指代消解性能对比

人称代词类型	*P*/%	*R*/%	*F*1/%
第一人称	79.54	66.03	72.16
第二人称	76.19	64.00	69.56
第三人称	88.37	73.07	80.00

由表7可知,第三人称指代消解性能最高,*F*1值达到了80%,第一人称指代消解性能次之,*F*1值为72.16%,第二人称指代消解性能最低,*F*1值为69.56%.实验出现上述结果,是因为第三人称指代场景丰富,可以指代人类、物体等,且维吾尔语第三人称代词无性别之分,可以指代男性或女性;而第二人称除了某些特定的使用环境外,大多出现在篇章对话中,其指代场景与第三人称相比并不是非常复杂.所以在语料分布中,第三人称分布最广,第一人称分布次之,第二人称分布最少,表2统计的语料中人称代词子类分布也印证了上述特点.第三人称在数据中包含数量较多,存在分布广泛的特征向量供模型挖掘深层隐含信息,所以指代消解性能最优;而对分布数目相对较少的第二人称,包含的特征向量较少,模型对其学习和刻画能力也相对较弱,所以第二人称指代消解性能与第一人称和第三人称相比,指代消解性能较弱.

6 结束语

指代消解研究有利于自然语言处理技术的发展,现有研究主要针对汉语和英语等大语种,而未考虑上下文隐含语义信息的影响,且对维吾尔语等少数民族语言的指代消解研究相对匮乏.针对以上不足,本文采用双向机制的Bi-LSTM挖掘照应语和候选先行语上下文语义层面特征,另外,根据维吾尔语人称代词指代现象提取24个hand-crafted层面特征,融合上述两类特征完成维吾尔语人称代词指代消解任务.通过与LSTM对比,验证了引入双向机制下Bi-LSTM学习到的上下文语义层面特征,有效提高了维吾尔语人称代词指代消解性能,而与浅层机器学习算法SVM、ANN作对比,则验证了基于深度学习机制的Bi-LSTM具备挖掘隐藏语义信息和复杂数据内部结构的优势.此外,本文还从word embedding维度对性能的影响、人称代词子类指代消解性能、hand-crafted层面特征对指代消解性能的影响等角度探索了本文方法的有效性.

参考文献

[1] ZELENKO D, AONE C, TIBBETTS J. Coreference resolution for information extraction[A]. Proceedings of the ACL Workshop on Reference Resolution and Its Applications[C]. Barcelona, Spain: ACL, 2004. 9 – 16.

[2] VAN DEEMTER K, KIBBLE R. On coreferring: coreference in MUC and related annotation schemes[J]. Computational Linguistics, 2000, 26(4): 629 – 637.

[3] MCCARTHY J F, LEHNERT W G. Using decision trees for coreference resolution[A]. Proceedings of Fourteenth International Joint Conference on Artificial Intelligence[C]. Montreal: IJCAI, 2000. 1050 – 1055.

[4] SOON W M, Ng H T, LIM D C Y. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguistics, 2001, 27(4): 521 – 544.

[5] NG V, CARDIE C. Improving machine learning approaches to coreference resolution[A]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C]. Philadelphia: ACL, 2002. 104 – 111.

[6] 杨勇,李艳翠,周国栋,等. 指代消解中距离特征的研究[J]. 中文信息学报,2008,22(5):39 – 44.
YANG Y, LI Y C, ZHOU G D, et al. Reaserch on distance information for anaphora resolution[J]. Journal of Chinese Information Processing, 2008, 22(5): 39 – 44. (in Chinese)

[7] 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.
KONG F, ZHOU G D. Pronoun resolution in English and Chinese languages based on tree kernel[J]. Journal of Software, 2012, 34(5): 1085 – 1099. (in Chinese)

[8] 周炫余,刘娟,等. 基于测度优化Laplacian SVM的中文指代消解方法[J]. 电子学报,2016,44(12):3064 – 3072.
ZHOU X Y, LIU J, et al. Chinese anaphora resolution based on metric-optimized Laplacian SVM[J]. Acta Electronica Sinica, 2016, 44(12): 3064 – 3072. (in Chinese)

[9] 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.
XI X F, ZHOU G D. Pronoun resolution based on deep learning[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 100 – 110. (in Chinese)

[10] 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.

[11] KOLEN J, KREMER S. Gradient flow in recurrent nets: the difficulty of learning longterm dependencies[J]. Wiley-IEEE Press, 2001, 28(2): 237 – 243.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1681 – 1726.

[13] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirec-

tional long short-term memory networks for relation classification[A]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics[C]. Berlin, Germany: ACL, 2016. 207 – 212.

[14] TANG D, QIN B, FENG X, et al. Effective LSTMs for target-dependent sentiment classification[A]. Proceedings of the 26th International Conference on Computational Linguistics [C]. Osaka, Japan: COLING, 2016. 3298 – 3307.

[15] BENGIO Y, DUCHARME R, JEAN, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3(6): 1137 – 1155.

[16] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[A]. Proceedings of the 25th International Conference on Machine Learning[C]. Helsinki, Finland: ICML, 2008. 160 – 167.

[17] 程适良. 现代维吾尔语语法[M]. 新疆: 新疆人民出版社, 1996.

[18] 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.

作者简介

田生伟　男, 1973 年出生, 新疆乌鲁木齐人, 博士, 现为新疆大学教授, 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

秦　越　女, 1992 年出生, 新疆昌吉人, 现为新疆大学硕士研究生, 主要研究领域为自然语言处理.
E-mail: qinyue_xju@163.com

禹　龙(通信作者)　女, 1974 年出生, 新疆乌鲁木齐人, 硕士, 现为新疆大学教授, 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