基于指代消解的民间文学文本实体关系抽取

魏静,岳昆,段亮,王笳辉

(云南大学 信息学院;云南省智能系统与计算重点实验室,昆明 650500)

摘 要:民间文学是中华文化的重要组成部分,具有重要的研究价值。随着人工智能的快速发展,数字化技术成为修复民间文学残缺作品、构建民间文学领域知识图谱等实际应用的重要方式.然而,民间文学文本中指示代词多、实体关系重叠,使得民间文学文本关系抽取困难.为此,提出一种基于指代消解的实体关系联合抽取方法 CR_RSAN,使用指代消解获取指示代词和对应实体的位置信息,并利用该信息设计指示代词替换算法和调整文本序列标注方法,以此强化模型获取文本语义特征的能力.此外,使用同时编码实体和关系信息的序列标注方法以缓解文本实体关系重叠问题.对比实验选用目前主流方法的模型作为基线,并相继在民间文学文本上进行实验,CR_RSAN 在精确率、召回率和 $F1$ 值等方面分别提高了 13.39 个百分点、14.29 个百分点和 14.98 个百分点.

关键词:民间文学;关系抽取;指代消解;注意力;序列标注

中图分类号:TP391 **文献标志码**:A **文章编号**:1000-2367(2024)01-0084-09

民间文学是指民众在生活文化和生活世界里传承、传播、共享的口头传统和语辞艺术,是中华文化的重要组成部分.民间文学包括神话、民间传说、民间故事、韵文的诗歌等体裁的民间作品.传统的民间文学神话诗歌多以文本为载体,或以口头传唱的方式讲述神话故事,不易开展后续的研究和保护工作.随着大数据和人工智能技术的快速发展和不断普及,数字媒介成为民间文学传播的重要载体,是民间文学文本保护和传承的重要方式[1-2].将民间文学文本与自然语言处理相关技术结合,以揭示神话故事蕴含的内在特征,建立对文本数据的整体认知,有助于研究者和民间文学爱好者对民间文学文本的了解和学习,为民间文学数字化提供了新型的技术支撑.

关系抽取(relation extraction,RE)是从文本中识别出实体,抽取实体间的语义关系,旨在将非结构化的文本数据转化为结构化数据[3].利用关系抽取对民间文学文本进行数字化处理,其处理结果可应用于民间文学文本中的诗歌修复,民间文学的领域知识图谱构建等实际应用.民间文学与新兴领域技术结合,增加民间文学的受众群体,有利于发扬和传承民间文学工作.此外,补全民间文学中诗歌的缺失部分,弥补了人员能力有限的缺陷,更大程度上确保了民间文学文化遗产的完整性.

民间文学文本以诗歌为载体,具有以下特点:(1)文本指示代词多,实体关系常由指示代词体现,在指示代词之间容易隐藏实体间的真实关系,是实体间关系识别的重要影响因素.(2)民间文学文本有着丰富的语义表达,在一段简单的文本描述中包含多个实体-关系三元组,且同一实体不只存在一种关系,可见如何处理关系重叠和实体重叠是民间文学文本关系抽取模型需要解决的问题.综上所述,民间文学文本有着指示代词

收稿日期:2022-06-23;**修回日期**:2022-09-05.

基金项目:云南省重大科技专项(202002AD080002);云南省教育厅科学研究基金(2002Y010);云南大学研究生科研创新项目(2021Y023;2021Y174).

作者简介:魏静(1998—),女,四川越西人,云南大学硕士研究生,研究方向为知识图谱、自然语言处理.

通信作者:岳昆(1979—),男,云南大学教授,博士,研究方向为海量数据分析、知识发现、贝叶斯深度学习,E-mail:kyue@ynu.edu.cn.

引用本文:魏静,岳昆,段亮,等.基于指代消解的民间文学文本实体关系抽取[J].河南师范大学学报(自然科学版),2024,52(1):84-92.(Wei Jing,Yue Kun,Duan Liang,et al.Coreference resolution for relation extraction in folk literature[J].Journal of Henan Normal University(Natural Science Edition),2024,52(1):84-92.DOI:10.16366/j.cnki.1000-2367.2022.06.23.0001.)

较多，实体重叠和关系重叠的特点，故面向民间文学文本的关系抽取模型应具有更善于获取文本语义特征，实体与关系之间的交互信息特征及处理和利用文本中指示代词信息的能力.

但目前的方法[4－15]或是偏向较为宽泛的领域或是面向如医学、生物学等特定领域.若直接将模型迁移到民间文学文本关系抽取任务中，不能真正解决民间文学文本代词多、实体关系重叠带来的问题.

根据上述民间文学文本特点及当前实体关系抽取方法的研究进展，本文提出一种基于指代消解的民间文学文本关系抽取模型 CR_RSAN.针对文本中指示代词较多、易造成实体关系不明的问题，引入指代消解[16]方法，通过不断学习文本中每个词语和可能存在的先行词的特征，计算文本中所有词的关联分数，得到指示代词及其对应实体的位置信息.在此基础上，设计指示代词替换算法，将指示代词和其对应实体进行替换，使实体关系明朗，帮助模型获取到更多的实体关系.

针对文本实体关系重叠的问题，本文采用序列标注的方式来获取实体间关系.经指示代词替换之后的文本比原文本有更强的实体关系表现力，故根据替换后文本中实体的位置分布信息，在 BIESO 标注方法结合新编码 H 和 T[9]的基础上，调整头实体和尾实体的编码位置，将指示代词替换后提供的有效信息与实体关系间的交互信息用序列标注的方式结合，提高模型获取语义特征的能力，将实体关系的重叠问题转化为序列标注问题，以提高模型关系抽取性能.

在民间文学文本上的实验结果表明，CR_RSAN 在精确率、召回率和 $F1$ 值上均优于现有模型.

1　相关工作

在早期的研究中，基于核函数的关系抽取方法，需要大量的人力来标注数据，且特征提取中存在误差，影响实体间关系的抽取效果[17].目前基于深度学习的实体关系抽取方法成为研究关系抽取的重要方式[18]，一些方法[19]将实体命名识别和关系抽取任务分开进行，造成实体识别任务出现的错误累积到关系抽取任务中，影响了关系抽取的效果[3].

为了缓解实体命名识别任务造成的错误累积，研究者们相继提出将实体命名识别和关系抽取联合进行的方法.KATIYAR 等[4]通过共享编码层特征表示联合提取实体和关系，但实体识别任务和关系抽取任务是单独解码，这导致实体与关系之间的交互信息易被遗漏.为了利用该信息，研究者们设计了不同的标记策略.ZHENG 等[5]提出采用端到端的模型直接对实体-关系三元组建模，但该模型无法解决关系重叠问题.ZENG 等[6]通过复制机制(copy mechanism)生成后续的两个对应实体的关系，解决了关系重叠的问题，但抽取结果不完整.WEI 等[7]提出二进制标记的方法解决相同实体对之间包含多重关系的问题，但对语义较为复杂的文本数据，该模型学习特征的能力有限.YAN 等[8]提出将神经元划分，但其划分规则忽视了实体特定区域中与关系特征的相关信息.YUAN 等[9]提出改进 BIESO 标注，将关系信息作为先验知识，减少模型对无关实体的关注，解决了实体关系重叠的问题，但忽视了某些关系发生的真正主体.

在医学领域，宁尚明等[10]采用多通道注意力机制与卷积神经网络相结合，但是将实体识别任务与关系抽取任务分开进行，忽略了两者之间的依赖.LI 等[11]采用共享双向长短记忆网络编码层，XUE 等[12]采用共享 BERT 模型[20]的编码，实现实体命名识别任务和关系抽取任务的参数共享，但两个任务分开解码忽略了实体与关系之间的交互信息.在文学领域，秦川等[21]利用主题、模板信息和押韵信息实现秘密信息的隐藏，YI 等[22]提出利用分解子空间进行对抗训练以区分不同风格的诗歌.

基于上述研究，现有的关系抽取模型侧重于研究如何利用实体信息、关系信息和实体与关系之间的交互信息以提高模型性能，但是抽取结果不尽如人意，特征信息不能完全利用，同时未对文本数据中的指示代词进行处理，若直接将这些方法迁移到民间文学文本关系抽取任务中，并不能真正解决指示代词带来的问题.LEE 等[16]提出一种端到端的指代消解方式，把文本中的所有词语视为潜在实体，通过学习指示代词与先行实体之间的相关性，确定指示代词与先行实体的位置关系，解决文本中指示代词与对应实体的指代问题.因此，针对民间文学文本指示代词多，实体关系重叠的特点，将指代消解和关系抽取相结合给民间文学文本关系抽取任务提供了新思路.

2 CR_RSAN 模型

2.1 问题描述

设民间文学文本为 D,实体-关系三元组表示为$(e_h,r_s,e_t)(e_h,e_t\in E,r_s\in R)$,$E$ 和R 分别表示实体集合和关系集合.

给定长度为 n 的文本 D 和一个预定义的关系集合R,模型的任务是利用指代消解得到文本 D 中的指示代词 m 和其对应实体m_h 的位置信息,根据位置信息替换文本中的指示代词得到新的文本$\widetilde{D}$,从$\widetilde{D}$ 中识别所有存在的三元组(e_h,r_s,e_t).

2.2 模型介绍

模型主要包括输入层、指代消解、指示代词替换,序列标注和实体关系联合抽取.

(1)输入层:原始的民间文学文本作为模型的输入.

(2)指代消解:对输入的文本数据进行词语划分,学习和计算每个词及其先行词的关联分数,识别文本数据中存在的指示代词及其对应实体.

(3)指示代词替换:按照指示代词替换算法对文本数据中的指示代词进行有效替换.

(4)序列标注:根据(3)中的替换结果制定新的序列标注方法,生成不同关系下的文本标注序列.

(5)实体关系联合抽取:利用基于关系的注意机制生成每个关系特定的文本表示,提取当前关系下存在的头尾实体.

以句子"九隆将身一纵跳上毒龙龙头,他双手举起长刀迎头就砍".为例,CR_RSAN 的总体架构图如图 1 所示.句子经过输入层进入指代消解部分,得到指示代词"他"和对应实体"九隆"的位置信息,利用该信息将文本中的指示代词进行替换生成新的文本,输入到实体关系联合抽取模型 RSAN 中,得到文本中包含的实体-关系三元组(九隆,敌对,毒龙)和(九隆,动作,长刀).

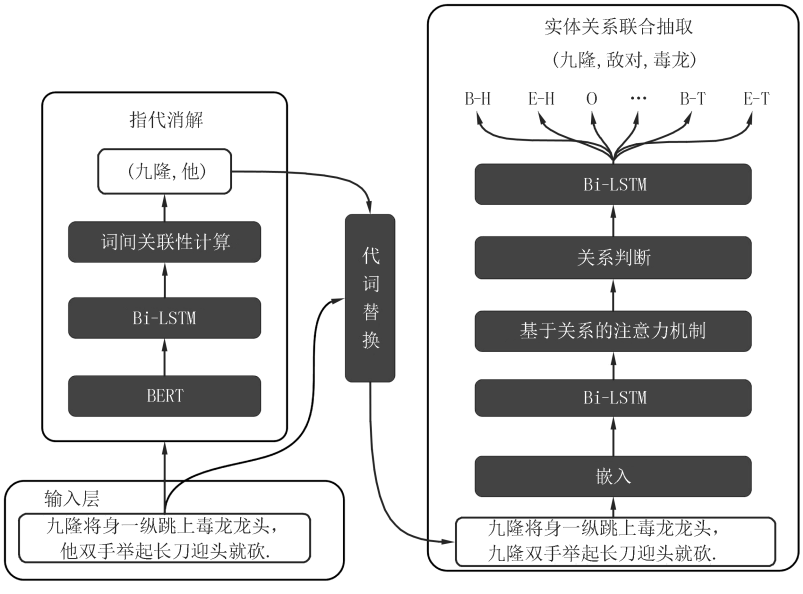

图1 模型架构图

Fig.1 Model architecture diagram

2.2.1 指代消解

民间文学文本中包含较多指示代词,而指示代词会影响实体关系确定,因此利用指代消解方法获取指示代词对应的实体信息,有助于后续任务发现隐藏的实体间关系.指代消解的基本原理是:将文本 D 进行分词,分词后的文本表示为 $D_p=\{p_1,p_2,\cdots,p_h\}(h\leq n)$,$p_i(1\leq i\leq h)$ 表示第 i 个位置上的词,计算任意两个词的关联分数(即代表相同事物的可能性),关联分数高的词语视为它们的指代相同,从而得到文本中的指示

代词与对应实体的位置信息.

2.2.1.1 词特征表示

在计算词之间的关联性时,首先需要得到词的特征表示.通过 BERT 模型[20]生成初始文本 D 的向量表示 $\boldsymbol{d}=\{\boldsymbol{d}_1,\cdots,\boldsymbol{d}_t,\cdots,\boldsymbol{d}_n\}$,$\boldsymbol{d}_t(1\leq t\leq n)$ 表示第 t 个位置上的字.把 $\boldsymbol{d}_t$ 作为 Bi-LSTM 网络的输入,获得具有上下文语义信息的文本表示 s_t,使用前反馈神经网络 F 将向量形式的 $\boldsymbol{s}_t$ 转化为其对应的数值形式 $\propto_t$,用于注意力训练得到词 p_i 的覆盖字向量 $\hat{\boldsymbol{s}}_i$,最后通过式(4)计算词 p_i 的特征表示.

$$\propto_t=F(\boldsymbol{s}_t),\tag{1}$$

$$a_{p_i,t}=\frac{\exp(\alpha_t)}{\sum_{k=START(p_i)}^{EDN(p_i)}\exp(\alpha_k)},\tag{2}$$

$$\hat{\boldsymbol{s}}_i=\sum_{t=START(p_i)}^{END(p_i)}a_{p,i,t}\cdot\boldsymbol{s}_t,\tag{3}$$

$$\boldsymbol{g}_{p_i}=[\boldsymbol{s}_{START(p_i)},\boldsymbol{s}_{END(p_i)},\hat{\boldsymbol{s}}_i,\rho(p_i)],\tag{4}$$

$START(p_i)$ 和 $END(p_i)$ 分别表示词 p_i 开始位置索引和结束位置索引,$\rho(p_i)$ 表示词 p_i 的长度大小.

2.2.1.2 关联分数计算

本文采用打分制来确定指示代词和其真实实体的位置信息.设 p_i 的可能先行词为 $p_j(1\leq j\leq i)$,p_i 和 p_j 之间的关联程度越强说明 p_j 是 p_i 的先行词的可能性越大.故采用关联分数来表示词之间的关联强度.

将词的向量表示 $\boldsymbol{g}_{p_i}$ 通过前馈神经网络 F_o 得到其对应的数值形式 $score_o(p_i)$ 用于后续计算;通过式(6)计算词 p_j 是词 p_i 先行词的分数 $score_a(p_i,p_j)$;通过式(7)计算词 p_i 和词 p_q 之间的关联分数 $score(p_i,p_j)$,即认为词 p_j 是词 p_i 的先行词,并且词 p_j 和词 p_i 同时存在才有关联的可能[16].通过计算词 p_i 与先行词的关联分数得到每个先行词与其关联的可能性,可能性最大的先行词视为词 p_i 的指代实体,最后得到文本中指示代词 m 和其对应实体 m_h 的位置索引.$\phi(p_i,p_j)$ 表示词 p_i 和词 p_j 之间的距离.

$$score_o(p_i)=\boldsymbol{W}_o\cdot F_o(\boldsymbol{g}_{p_i}),\tag{5}$$

$$score_a(p_i,p_j)=\boldsymbol{W}_a\cdot F_a([\boldsymbol{g}_{p_i},\boldsymbol{g}_{p_j},\boldsymbol{g}_{p_i}\circ\boldsymbol{g}_{p_j}],\phi(p_i,p_j)),\tag{6}$$

$$score(p_i,p_j)=score_o(p_i)+score_o(p_j)+score_a(p_i,p_j).\tag{7}$$

2.2.2 指示代词替换算法

指示代词使实体间关系不明显,若能将文本中的指示代词换成其对应实体,能帮助模型获取更确切的语义信息,提高实体关系抽取任务的性能.通过指代消解获取文本中指示代词与其对应实体的位置关系,利用该信息对指示代词和其对应实体进行有效替换.

本文将民间文学文本中出现的指示代词分为有效指示代词和无效指示代词.若文本数据中指示代词对应的实体是同一文本中实体-关系三元组中的头实体或者尾实体,则表示该指示代词与实体关系的确定有影响,将该指示代词视为有效指示代词;若文本数据中指示代词对应的实体不是实体-三元组中的头实体或者尾实体,则该指示代词视为无效指示代词.

如图 2 所示,情况二中实体间关系通过指示代词所在的短句体现和表达,将该指示代词替换成其对应实体,能增强实体与实体间关系的表现,有利于实体间关系抽取.

根据指示代词的分类,本文制定以下替换规则:若文本 D 中代词 m 对应的实体 m_h 是民间文学文本中关系三元组 (e_h,r_s,e_t) 包含的头实体或尾实体,则将文本 D 中的代词替换为对应实体,否则对文本不作处理.根据不同的实体关系同一个句子采取不同的替换策略,算法 1 给出了指示代词替换的具体步骤.

算法 1　指示代词替换

输入:　文本 D,(e_h,r_s,e_t),(m,m_h)

输出:　$\widetilde{D}$

1. IF $m_h=e_h$ OR $m_h=e_{\text{tail}}$ THEN
2.　IF $m_h=e_h$ THEN
3.　$\widetilde{D}\leftarrow D$ 中的 m 替换为 e_h
4.　END IF
5.　IF $m_h=e_t$ THEN
6.　$\widetilde{D}\leftarrow D$ 中的 m 替换为 e_t
7.　END IF
8. ELSE $\widetilde{D}\leftarrow D$

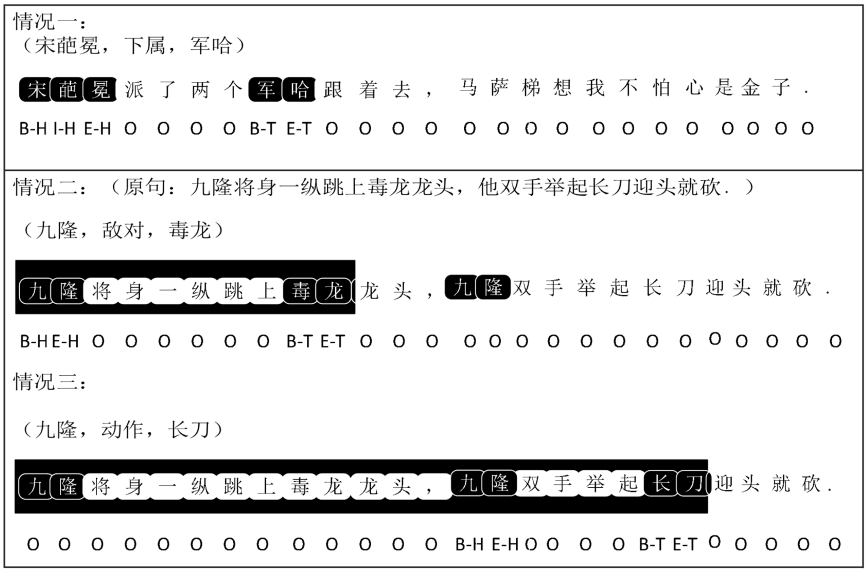

图2 代词替换和序列标注示例

Fig.2 The sample of demonstrative pronoun replacement and sequence labeling

2.2.3 序列标注

YUAN 等[9]将{H,T}并入 BISO 标注中作为实体标注的后缀，分别表示头实体和尾实体，与实体无关的用 O 表示.针对民间文学文本代词多、关系重叠的特点，本文在标注方法上做以下改进：

(1)未经过指示代词替换的文本，则对文本数据首次出现的头实体和尾实体进行 BISO 实体标注，并分别以 H 和 T 作为后缀，其他与首尾实体无关的部分用 O 表示.

(2)经过指示代词替换的文本数据分为以下两种情况：①若指示代词替换成对应实体的位置位于替换前文本中头实体首次出现和尾实体首次出现的位置之间，则对指示代词替换后的实体和文本中首次出现的尾实体进行 BISO 实体标注，并分别以 H 和 T 作为后缀，剩余部分用 O 表示；②若指示代词替换成对应实体的位置不在替换前文本中头实体首次出现和尾实体首次出现的位置之间，则对文本数据首次出现的头实体和尾实体进行 BISO 实体标注，并分别以 H 和 T 作为后缀，其他与首尾实体无关的部分用 O 表示.图 2 为本文制定的序列标注方法的示例.

在文本关系抽取中，关系存在才使得实体存在是有意义，且在不同的关系中，字词对文本的潜在语义表达有着不同的影响，这使得字词在不同关系下的重要程度有所不同.故基于上述提出的序列标注方法，本文使用 RSAN 模型[9]利用基于关系的注意力机制为不同关系下的文本 $\widetilde{D}$ 生成不同的标注序列 S，并通过式(9)对标注序列进行解码，得到文本中的实体-关系三元组(e_h,r,e_t)，

$$S=RSAN(\widetilde{D}), \tag{8}$$

$$(e_h,r,e_t)=DECODE(S). \tag{9}$$

3 实 验

3.1 数据集

本文实验使用的数据集来自云南大学文学院提供的《千瓣莲花》《娥并与桑落》《傣族民间故事选》和《云南少数民族古典史诗》等书籍资料，共计 30 万字左右，指示代词占比约为 4.28%，整理得到 1 154 条数据.CR_RSAN的训练部分包括了指代消解和关系抽取，故数据集应包含两部分的模型训练输入，数据集中每条数据采用如图 3 中实例的标注方法，其中，cluster 表示文本中指示代词的位置索引，subtoken_map 表示文本词语划分情况.

将数据集按照 6∶2∶2 划分为训练集、验证集和测试集.数据集中包含 18 种关系，每种关系对应的数量

情况如表 1 所示.


```
"sentText": "岩罕接到命令觉得脖子上已架上凉冰冰的刀，他急忙跑回家找妻子商量．",
"relationMentions": [{"em1Text": "岩罕", "em2Text": "妻子", "label": "夫妻"}],
"entityMentions": ["妻子", "岩罕"],
"sentences": [["岩", "罕", "接", "到", "命", "令", "觉", "得", "脖", "子", "上", "已", "架", "上",
               "凉", "冰", "冰", "的", "刀", "，", "他", "急", "忙", "跑", "回", "家", "找", "妻",
               "子", "商", "量", "．"]],
"clusters": [[[0, 1], [20, 20]]],
"speaker_ids": [["a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a", "a",
                 "a", "a", "b", "b", "b", "b", "b", "b", "b", "b", "b", "b", "b", "b"]],
"sentence_map": [[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                  1]],
"subtoken_map": [[1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 8, 8, 9, 9, 9, 10, 11, 12, 13, 14, 14, 15, 16, 16,
                  17, 18, 18, 19, 19, 20]],
"genre": "dummy_genre",
"doc_key": "dummy_data"
```


图3　数据标注示例

Fig. 3　The sample of data labeling

表 1　关系类型及数量情况

Tab. 1　The type and quality of relations

关系类别	演奏	坐骑	爱慕	晚辈	住所	变体	制作	去处	夫妻
关系数量	21	56	68	248	97	36	38	218	85
关系类别	下属	继承人	称谓	兄弟	配饰	拥有	姐妹	敌对	朋友
关系数量	250	9	282	75	55	234	26	223	163

3.2　对比方法

本文选取 RSAN，Bias，PFN 和 CasRel 模型与 CR_RSAN 进行性能比较.

Bias[5]：将 BIESO 标注与{1，2}(1，2 分别表示实体 1 和实体 2)相结合生成新标签序列，并将联合抽取模型转换为一个序列标注问题的方法.

CasRel[7]：提出采用联级二进制的方法标注数据，将关系抽取任务从关系分类转换为主体向客体映射的方法.

PFN[8]：将神经元分割成两个任务分区(实体分区和关系分区)和一个共享分区，任务分区相互独立，共享分区保证实体信息和关系信息的交互的方法.

RSAN[9]：一种新型标注方式，将头实体和尾实体信息{H，T}标注进标签中，保留实体与关系之间的交互信息的方法.

3.3　评价指标

本文采用精确率(Precision，P)，召回率(Recall，R)和 $F1$ 值作为方法模型的评价指标.用 TP 表示预测为真，实际为真的样本数；FP 表示预测为真，实际为假的样本数；FN 表示预测为假，实际为真的样本数；TN 表示测试为假，实际为假的样本数.

精确率：表示在方法模型预测为真的样本中实际为真的比例，$P=\dfrac{TP}{TP+FP}$.

召回率：在实体为真的样本中，方法模型预测为真的比例，$R=\dfrac{TP}{TP+FN}$.

$F1$：是精确率和召回率的调和平均值，$F1=\dfrac{2PR}{P+R}$.

3.4　实验环境

实验平台使用 Windows 10 操作系统，配置为 8 核 Intel i7-10700K 处理器与 32 GB 内存，NVIDIA RTX 3070 显卡.所有模型基于 Python 和 Pytorch 构建.

3.5　实验结果

为了验证 CR_RSAN 的有效性，将 CR_RSAN 与 RSAN，PFN，Bias 和 CasRel 模型进行性能比较，其对

比结果如表 2 所示.

由表 2 可知,CR_RSAN 引入指代消解,将文本数据中的指示代词根据相应算法进行替换,加强模型获取语义特征的能力以及含有当前关系类型的实体对之间的联系,使性能显著提升,与 RSAN 模型相比精确率提高了 2.96 个百分点,召回率提高了 1.69 个百分点,$F1$ 值提高了 2.31 个百分点;与 Bias 模型相比精确率提高了 6.61 个百分点,召回率提高了 7.34 个百分点,$F1$ 值提高了 7.91 个百分点;与 CasRel 模型相比精确率提高了 12.06 个百分点,召回率提高了 12.29 个百分点,$F1$ 值提高了 12.54 个百分点;与 PFN 模型相比精确率提高了 13.39 个百分点,召回率提高了 14.29 个百分点,$F1$ 值提高了 14.98 个百分点.以上实验结果,验证了 CR_RSAN 模型的有效性.

表 2 模型的性能对比

Tab. 2 The comparison of model's performance

模型	民间文学文本			模型	民间文学文本		
	精确率/%	召回率/%	$F1$/%		精确率/%	召回率/%	$F1$/%
PFN	6.78	4.87	4.67	RSAN	17.21	17.47	17.34
CasRel	8.11	6.87	7.11	CR_RSAN	**20.17**	**19.16**	**19.65**
Bias	13.56	11.82	11.74				

本文经指代消解之后,采用改进的序列标注方法对文本标注,为了验证改进后的标注方法的有效性,现将采用原有标注方法 rr_origin 和 CR_RSAN 模型进行比较,其对比结果如表 3 所示.

由表 3 可知,改进后的序列标注方法结合了指示代词和对应实体之间的替换信息,使产生关系的两个实体能被标注,帮助模型获取实体间更明确的和贴切的语义信息.与采用原有标注方法的模型相比,采用改进后的序列标注方法的精确率提高了 0.6 个百分点,召回率提高了 0.41 个百分点,$F1$ 值提高了 1.31 个百分点.由此可见,改进后的序列标注方法对模型性能的提升作用.

表 3 序列标注方法对比

Tab. 3 The comparison of sequence labeling method

模型	民间文学文本		
	精确率/%	召回率/%	$F1$/%
rr_origin	19.57	18.75	18.34
CR_RSAN	**20.17**	**19.16**	**19.65**

3.6 民间文学文本实体关系抽取样例展示

对民间文学文本关系抽取可以快速了解文本中人物、事件之间的联系,方便研究者和民间文学爱好者快速建立对文学作品的整体认知.使用 CR_RSAN 和不加指代消解的方法 RSAN 分别对民间文学文本进行关系抽取,以此展示两种方法在关系抽取过程中的差别,如图 4 所示.

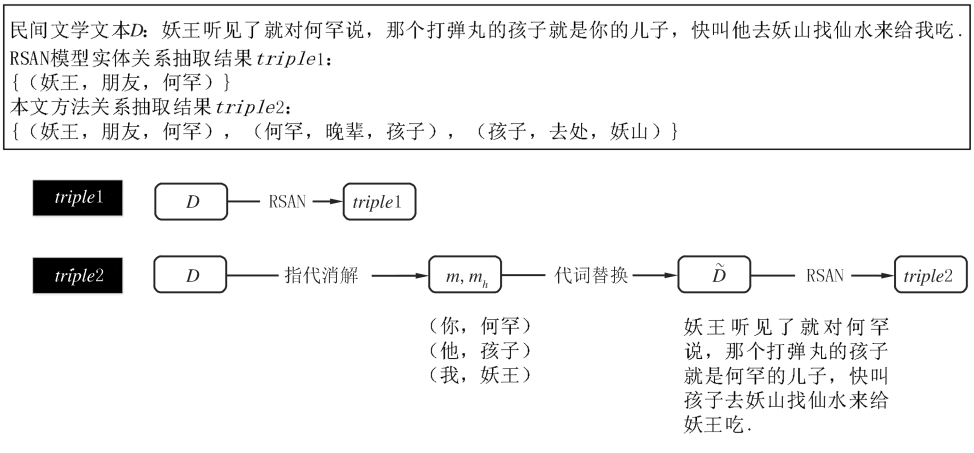

图4 民间文学文本关系抽取样例

Fig. 4 The sample of relation extraction in folk literature

图 4 中展示了民间文学文本 D 经 RSAN 模型生成了实体-关系三元组 $triple1$ 和经 CR_RSAN 模型生

成了实体-关系三元组 $triple2$.每个圆角矩形代表模型中关键节点的输出内容，m 代表文本中的指示代词，m_h 代表 m 的对应实体，$\tilde{D}$ 代表替换指示代词后的新文本.如例句经过指代消解后得到文本中的指示代词和对应的实体(你，何罕)、(他，孩子)、(我，妖王).将文本中指示代词替换之后生成新文本，再对新生成的文本进行关系抽取，得到文本中包含的实体-关系三元组 $triple2$.可见通过指代消解之后的关系抽取能获取更多的实体关系.

4　结　论

本文提出一种基于指代消解的关系抽取模型 CR_RSAN，用于民间文学文本的关系抽取任务.通过指代消解，获得指示词和对应实体的位置信息并对文本中的指示代词进行替换，并在 RSAN 模型标注方式上进行调整，将实体-关系抽取转换为标注问题.实验证明，模型在精确率、召回率和 $F1$ 值方面都有显著提升，这也证明了 CR_RSAN 在民间文学文本关系抽取任务中的有效性.

但是模型的精确率，召回率和 $F1$ 值均还有待提高，通过实验发现，该模型能很好地判断文本中含有什么关系，但是在抽取满足当前关系的实体对时会出现多字和少字的情况，因此降低了模型的精确率.在后续工作中，继续从数据集和模型优化着手，设计质量较高的数据集以及探索其他模型，学习其他模型的优点来改进模型，提高模型性能.

参　考　文　献

[1] AIDEN E L.Culturomics：Quantitative analysis of culture using millions of digitized books[C]//AIDEN Erez Lieberman.Culturomics：Quantitative analysis of culture using millions of digitized books.New York：Science，2011：176-182.

[2] 赵海英，贾耕云，潘志庚.文化计算方法与应用综述[J].计算机系统应用，2016，25(6)：1-8.
ZHAO H Y，JIA G Y，PAN Z G.Review on the methods and applications in cultural computing[J].Computer Systems & Applications，2016，25(6)：1-8.

[3] 鄂海红，张文静，肖思琪，等.深度学习实体关系抽取研究综述[J].软件学报，2019，30(6)：1793-1818.
HAIHONG E，ZHANG W J，XIAO S Q，et al.Survey of entity relationship extraction based on deep learning[J].Journal of Software，2019，30(6)：1793-1818.

[4] KATIYAR A，CARDIE C.Going out on a limb：joint extraction of entity mentions and relations without dependency trees[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.[s.l.：s.n.]，2017.

[5] ZHENG S C，WANG F，BAO H Y，et al.Joint extraction of entities and relations based on a novel tagging scheme[EB/OL].[2022-05-16].https：//blog.csdn.net/MaybeForever/ article/details/102668087.

[6] ZENG X R，ZENG D J，HE S Z，et al.Extracting relational facts by an end-to-end neural model with copy mechanism[EB/OL].[2022-06-21].https：//blog.csdn.net/Jeaksun/article/details/125350626.

[7] WEI Z P，SU J L，WANG Y，et al.A novel cascade binary tagging framework for relational triple extraction[EB/OL].[2022-05-13].https：//arxiv.org/abs/1909.03227.pdf.

[8] YAN Z H，ZHANG C，FU J L，et al.A partition filter network for joint entity and relation extraction[EB/OL].[2022-06-20].https：//zhuanlan.zhihu.com/p/559161506.

[9] YUAN Y，ZHOU X F，PAN S R，et al.A relation-specific attention network for joint entity and relation extraction[C]// IJCAI-PRICAI-20.[s.l.：s.n.]，2020.

[10] 宁尚明，滕飞，李天瑞.基于多通道自注意力机制的电子病历实体关系抽取[J].计算机学报，2020，43(5)：916-929.
NING S M，TENG F，LI T R.Multi-channel self-attention mechanism for relation extraction in clinical records[J].Chinese Journal of Computers，2020，43(5)：916-929.

[11] LI F，ZHANG M S，FU G H，et al.A neural joint model for entity and relation extraction from biomedical text[J].BMC Bioinformatics，2017，18(1)：198.

[12] XUE K，ZHOU Y M，MA Z Y，et al.Fine-tuning BERT for joint entity and relation extraction in Chinese medical text[C]//2019 IEEE International Conference on Bioinformatics and Biomedicine(BIBM).San Diego：IEEE，2020：892-897.

[13] SMIRNOVA A，CUDRé-MAUROUX P.Relation extraction using distant supervision：a survey[J].ACM Comput Surv，2019，51(5)：106.1-106.35.

[14] DAI D，XIAO X Y，LYU Y J，et al.Joint extraction of entities and overlapping relations using position-attentive sequence labeling[EB/OL].[2022-06-20].https：//blog.csdn.net/weixin_44729115/article/details/109594927.

[15] LIN Y K,SHEN S Q,LIU Z Y,et al.Neural relation extraction with selective attention over instances[C]// ACL2016.[s.l.:s.n.],2016.
[16] LEE K,HE L H,LEWIS M,et al.End-to-end neural coreference resolution[EB/OL].[2022-06-17].https://arxiv.org/abs/1707.07045.pdf.
[17] 常耀成,张宇翔,王红,等.特征驱动的关键词提取算法综述[J].软件学报,2018,29(7):2046-2070.
CHANG Y C,ZHANG Y X,WANG H,et al.Features oriented survey of state-of-the-art keyphrase extraction algorithms[J].Journal of Software,2018,29(7):2046-2070.
[18] KUMAR Shantanu.A survey of deep learning methods for relation extraction[EB/OL].[2022-06-18].https://www.doc88.com/p-5307498666672.html.
[19] WANG L L,CAO Z,DE MELO G,et al.Relation classification via multi-level attention CNNs[C]// ACL2016.[s.l.:s.n.],2016.
[20] DEVLIN J,CHANG M W,LEE K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[EB/OL].[2022-06-17].]https://arxiv.org/abs/1810.04805.pdf.
[21] 秦川,王萌,司广文,等.基于绝句生成的构造式信息隐藏算法[J].计算机学报,2021,44(4):773-785.
QIN C,WANG M,SI G W,et al.Constructive information hiding with Chinese quatrain generation[J].Chinese Journal of Computers,2021,44(4):773-785.
[22] YI X Y,LI R Y,YANG C,et al.MixPoet:diverse poetry generation via learning controllable mixed latent space[EB/OL].[2022-06-17].https://arxiv.org/abs/2003.06094.pdf.

Coreference resolution for relation extraction in folk literature

Wei Jing, Yue Kun, Duan Liang, Wang Jiahui

(School of Information Science and Engineering; Key Lab of Intelligent Systems and Computing of Yunnan Province, Yunnan University, Kunming 650500, China)

Abstract: Folk literature is an important part of Chinese culture and has significant value. With the rapid development of artificial intelligence, digital technology has become an important way to repair broken works and built the knowledge graph of folk literature. However, there are many demonstrative pronouns and overlapping entity relations in folk literature texts, which poses great challenges for the relation extraction in folk literature text. In view of these characteristics, the method named CR_RSAN is proposed for relation extraction which is based on coreference resolution. This method uses coreference resolution to obtain the position correspondence between demonstrative pronoun and the corresponding entity, and uses this information to design the demonstrative pronoun replacement algorithm and adjust the sequence labeling method, thus improving the ability of the method to obtain the semantic features of text. In addition, the sequence labeling method encodes both entity and relation information alleviates the problem of overlapping entity relation in text. Some methods with the best performance are selected as the baseline and verified in the folk literature text. CR_RSAN's precision, recall rate and $F1$ value are increased by13.39, 14.29 and 14.98 percentage points respectively.

Keywords: folk literature; relation extraction; coreference resolution; attention; sequence tagging

[责任编校 陈留院 赵晓华]