山西大学学报(自然科学版)48(3):470—480,2025
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024026

基于阅读理解的文档级人物属性抽取研究

刘资蕴,张世奇,陈文亮*
(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘　要:人物属性抽取旨在从人物介绍中抽取人物的各项属性,如性别、国籍等。已有抽取方法通常由序列标注模型对远程监督数据进行训练从而得到抽取模型,但是用该方式在数据上存在标注不准确和不同属性值重合的问题,在模型上缺少可扩展性和可泛化性能力。为解决此问题,该文提出将任务转化为阅读理解问题,通过阅读人物介绍来对人物属性表进行填写补全。为此,本文构造了一份基于阅读理解的文档级人物属性抽取数据集,并采用了基于Transformer算法的双向编码表征模型-机器阅读理解(BERT-MRC)和基于Transformer算法的双向编码表征模型-条件随机场-机器阅读理解(BERT-CRF-MRC)两种基线模型。研究结果表明BERT-CRF-MRC相比于BERT-MRC在*F*1值上高三个百分点,BERT-CRF-MRC的试验结果在短文本人物介绍中*F*1平均值约为92%,在长文本人物介绍中*F*1平均值约为75%。本文的新构建数据和代码已公开在Github上。

关键词:属性抽取;机器阅读理解;标注数据
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:0253-2395(2025)03-0470-11

Machine Reading Comprehension for Document-level Person Aspect Term Extraction

LIU Ziyun, ZHANG Shiqi, CHEN Wenliang*
(*School of Computer Science and Technology, Soochow University, Suzhou* 215006*, China*)

Abstract: Person aspect term extraction aims to extract various attributes of individuals such as gender and nationality from their descriptions. Existing extraction methods typically train sequence labeling models on distantly-supervised data to obtain the extraction model. However, this approach has issues with inaccurate annotations and overlapping different attribute values in the data, and lacks scalability and generalizability in their models. To solve the problems, this article proposes to transform this task into a machine reading comprehension (MRC) problem, that is, to fill in the person attribute-value table by reading the person profile. This paper constructs a person attribute recognition data based on the reading comprehension framework from the person encyclopedia, and constructs two baseline models of bidirectional encoder representations from transformers-machine reading comprehension (BERT-MRC) and bidirectional encoder representations from transformers-conditional random field-machine reading comprehension (BERT-CRF-MRC). Among them, BERT-CRF-MRC is three percentage points higher than BERT-MRC on average in *F*1 score and the experimental results of BERT-CRF-MRC are about 92% *F*1 average in short text person profiles while about 75% in long text person profiles. The constructed data and code are exposed on Github.

Key words: aspect term extraction; MRC; annotated data

收稿日期:2023-10-15;**接受日期**:2023-12-15
基金项目:国家自然科学基金(62376177)
作者简介:刘资蕴(1999—),男,内蒙古兴安盟乌兰浩特市人,硕士研究生,研究方向为知识图谱构建。E-mail:zyliu129@stu.suda.edu.cn
* **通信作者**:陈文亮(CHEN Wenliang),E-mail:wlchen@suda.edu.cn
引文格式:刘资蕴,张世奇,陈文亮.基于阅读理解的文档级人物属性抽取研究[J].山西大学学报(自然科学版),2025,48(3):470-480. DOI:10.13451/j.sxu.ns.2024026.

0 引言

属性抽取任务旨在从给定的半结构化文本或非结构化文本中抽取特定实体的三元组信息“<实体,属性,属性值>”[1],其中属性被预先定义并保存在属性列表中,属性值是从文本中抽取到的。表1展示了人物属性抽取的例子。

表1 人物“姚明”的三元组抽取结果

Table 1 The triple extraction result of the character "Yao Ming"

文本	姚明(Yao Ming),男,汉族,无党派人士,1980年9月12日出生于上海市徐汇区……
三元组	<姚明,性别,男> <姚明,民族,汉> <姚明,出生地,上海市徐汇区>

属性抽取任务在信息抽取任务中占据重要的地位。属性抽取的应用十分广泛,它是知识图谱构建的基础步骤之一,在电商领域中,属性抽取可以从海量的商品中找到有效商品信息从而进行检索、识别、判断两个商品是否相同[2],给客户带来更好的购物体验;在医疗领域中,属性抽取可以从海量的病例中抽取疾病的相关症状,从而更好地辅助医生进行治疗[3]。

面向属性抽取任务构建的评测数据集来源多种多样,通用数据为国际语义评测会议(SemEval)的4个基准数据集[4],该数据集不仅提供了商品类的属性标注结果,也标注了每个属性在句子内的情感词性。Hu等[5]提出了CRD(Customer Review Dataset)数据集,包含了5种电子产品的用户评论,也是最早应用于属性抽取任务的数据集之一。此外还有很多类似的数据集,但由于属性抽取任务差异较大,构建一个通用的大规模数据集是不切实际的,因此研究者通常会基于特定的领域构造自己的数据进行研究。为缓解人工标注耗时耗力的问题,研究者通常会采用远程监督的方式从大量非结构文本中进行标注。本文聚焦于人物属性抽取,从百科数据中抽取海量人物数据用于构建人物属性抽取语料,已经被证明是一种可行的方案[6]。

属性抽取的关键技术在逐渐发展。早期主要采用基于规则的属性抽取,但随后逐渐减少,研究重心转移到基于神经网络的深度学习技术。相比基于卷积的语义编码方法,属性抽取研究更倾向于采用基于序列的循环计算编码技术,其中以集成了长短期记忆单元(Long-term Short Memory,LSTM)[7]的循环神经网络最为常见。这是因为属性抽取任务的处理对象是自然语言中的词语,对任意属性判别都强烈依赖于上下文信息。LSTM能够在一定程度上缓解长距离依赖问题,即通过融合上下文语义信息来改善记忆处理,因此成为近期研究中不可或缺的技术之一。

条件随机场(Conditional Random Field,CRF)[8],作为一种基于动态规划的选择方法,因其在序列标注场景中的天然适用性而被广泛应用。特别是,在进行最优化判别过程中,CRF依赖于之前时刻的最优解,这种特性使其与LSTM具有较高的兼容性。例如,2019年提出的前沿模型面向目标的意见词提取模型(Target-oriented Opinion Words Extraction,TOWE)[9]和结合规则的神经网络属性抽取模型(Rule Incorporated Neural Aspect and Opinion Term Extraction,RINANTE)[10]便是结合了CRF和LSTM进行建模的典型案例。

此外,注意力机制(attention)[11]也是属性抽取领域频繁采用的一项关键技术,尤其是在识别和加权上下文中的高关注度信息方面发挥着重要作用。值得注意的是,集成了多层多头注意力机制的Transformer[11]架构可以视为预训练语言模型在属性抽取任务中成功应用的一个例证,其中最为突出的就是基于Transformer算法的双向编码表征模型(Bidirectional Encoder Representations from Transformers,BERT)[12]预训练语言模型,它使用双向掩码(MASK)机制来保证了语义前后的通顺性。属性抽取模型目前主要有两种成功范式,一种是以BiLSTM-CRF[13]为代表的序列标注模型,另一种是以基于Transformer算法的双向编码表征模型-机器阅读理解(Bidirectional Encoder Representations from Transformers-machine Reading Comprehension,BERT-MRC)[14]为代表的阅读理解模型。目前大都以序列标注模型进行人物属性抽取[15]。

本文通过观察数据人物属性数据,发现序列标注模型在标注数据时会产生如下问题:

(1)标注错误:在文本中出现与属性值相同字符的样例,但它并不是属性值。例如在文本“马克·斯坦恩,著有《衰亡的美国》”中,该人物的国籍是美国,但是文中的美国是人物的作品名字,不能被标注为国籍,如果按序列标注模式的标注方法,则会被标记。这么做会导致模型学到错误的属性信息。

(2)不同属性的属性值重合问题:两个属性的属性值是同一个的话,依照序列标注模型则只能选择一个。这会使模型学偏,倾向于标注的属性。例如“出生地”属性与“国籍”属性很有可能是相同的。这样会减少标注的数据量,同时也会使模型偏向于总是抢先得到标注的属性。

基于上述问题,本文基于机器阅读理解方法,从海量百科数据中构建一份文档级别的人物属性抽取数据。该数据形式可表示为一段人物介绍加上这个人的属性表。如图1所示,本文让模型阅读这份人物介绍,然后输出该人物各个属性的属性值,即填充人物属性表。通过对比已经标注的属性表和模型输出的属性表,就可以知道模型的好坏。

对于第一种的标注错误,因为无法知道究竟文本中的哪个“美国”才是真正代表国籍的

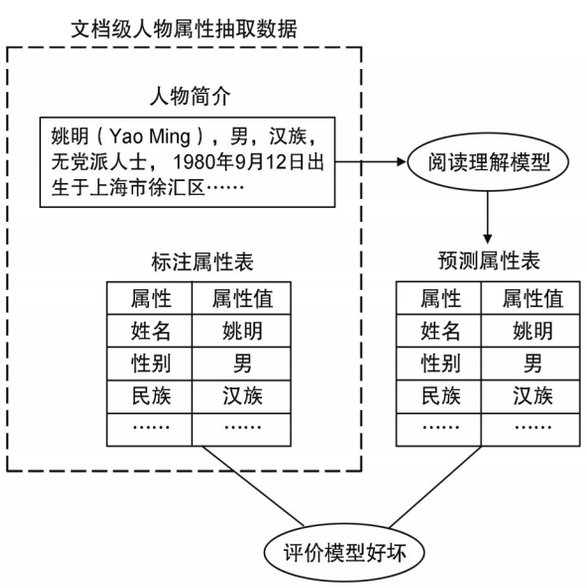

图1 文档级人物属性抽取数据形式样例

Fig. 1 Sample data format for extracting document level character attributes

属性值,所以不妨把它单独拿出来,把属性值保存在属性表中,而不是直接标注在文本中,因此不会产生直接用序列标注方法而产生的标注错误。但与此同时也失去了文中的标注信息,这需要模型去解决。

对于第二种的重合问题,重合的属性值都可以放在属性表中,因此也解决了序列标注中因属性值重合无法标注的情况。

数据构建框架如图2所示,本文在构建数据的过程中,做出了如下举措,最终抽取出44万条数据。

(1)为了让属性表中的数据更加合理,尽量避免出现由于属性表本身错误导致的错标、漏标等情况,本文对大部分属性值做了标准化处理。

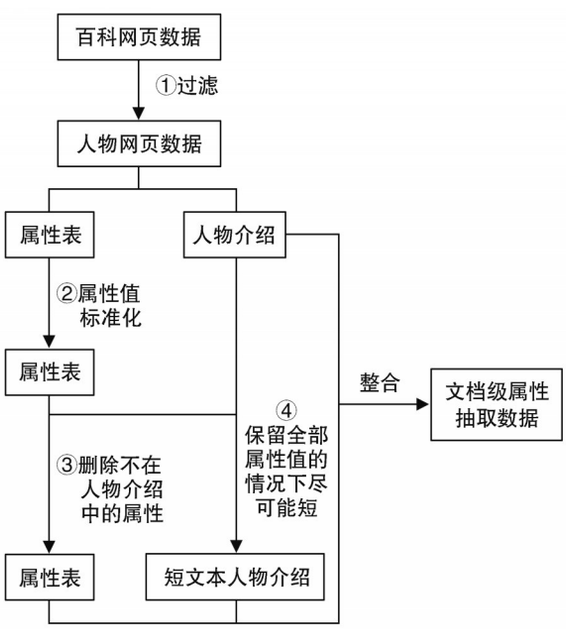

图2 数据构建框架

Fig. 2 Framework for data construction

(2)确保标注在属性表中的属性值一定出现在人物介绍中。本文把人物介绍分为两种,一种是长文本人物介绍,即全部有关该人物的信息,另一种是短文本人物介绍,该文本尽可能得短,并且同样包含属性表中的全部属性值。这么做的原因是要获取高质量的人物介绍,并分析模型在这两种数据下的效果。

(3)考虑人物属性的复杂度和出现的频率,最终选取了16种人物属性。

最后,本文使用了问答模型范式,用

BERT-MRC和BERT-CRF-MRC两个基线模型构建属性抽取系统,并对这两份数据进行了实验。

综上,本文的贡献如下:

(1)指出了远程监督应用于序列标注模型的标注局限,构建了大规模基于阅读理解的文档级别的人物属性抽取数据集。数据包括长文本和短文本(内容尽可能短,但包含属性表中全部属性的属性值)两种类型。构建的数据代码公开在https://github.com/ReturnTR/DPAR上。

(2)构建了BERT-MRC和BERT-CRF-MRC两个基线模型,对构建的数据集训练和评测。

1 相关工作

这里主要讨论针对人物的属性抽取相关工作。人物属性抽取任务并没有一个标准的公开数据集供研究者研究。通常情况是基于任务的内容来构建数据。李红亮[6]从百科网页中按规则抽取出人物介绍的文本,用触发词的规则方法对该数据进行抽取实验。张巧等[16]从导师个人主页上获取810张共九种属性的英文语料库,采用弱监督的方法提取属性中的模式用于抽取属性。Angeli等[17]基于主动学习和远程监督构建了一些高质量的人物属性标注数据。马进等[15]采用远程监督的方式从百科数据中抽取属性表和人物介绍,然后将其应用在序列标注模型上,得到了很好的效果。

早期的人物属性抽取方法为基于规则的方法[6]。该方法针对的属性模式较为简单,由专家设计特定的抽取模式规则,然后用这些规则在文中匹配出属性和属性值。例如“周杰伦,男,台湾出生”这句话中“,男,”可以作为识别性别属性的一个规则。该方法抽取的好坏取决于规则设计的是否完善合理。

随着属性抽取涉及的领域越来越多,为各个领域设计不同的规则会显得费时费力,这就出现了能够适应不同领域的属性抽取模型,即传统的机器学习模型[18]。该模型利用句子本身存在的语法,句法的关系提取句子中的特征,典型的模型如条件随机场CRF[19],该模型通过设计特征模板来提取出句子的特征,然后学习如何利用这些特征来预测出属性值。

近年来,由于深度学习模型强大的学习能力,研究者们逐渐采用深度学习模型自动学习出特征代替手动设计特征模板。例如LSTM[20]和门控网络(Gated Recurrent Unit,GRU)[21]。但是这些方案仍需要大量的标注数据用于训练模型,事实上,大部分自然语言处理任务都受限于标注数据的大小。受此限制,一种基于迁移学习的训练模式更受欢迎,该方法在其他的大量数据集中训练自然语言处理(Natural Language Processing,NLP)的简单任务,然后在新的任务中实现参数共享,训练时进行微调,使得新任务不再是从头开始训练。这类模型称为预训练模型。例如生成式预训练模型(Generative Pretrained Transformer,GPT)[22]、ELMO(Embeddings from Language Models)[23],以BERT[12]为基础的预训练模型效果最佳。

本文在马进等的百科信息抽取工作[15]基础上,将抽取问题转化为阅读理解问题,并为这种类型的问题设计了相应的数据格式,从而构建文档级别的属性抽取数据。

2 文档级人物属性标注语料库构建

2.1 人物数据获取

数据来自百度百科HTML界面,基于任务的需要,本文从网页中抽取出四项:名称、简介、属性表以及其他有关该实体的信息,以篮球运动员“姚明”为例,抽取的部分结果如表2所示。

通过观察网页规则可以将这四项抽取出

表2 人物“姚明”的四项抽取结果

Table 2 The four item extraction results of the character "Yao Ming"

名称	姚明(亚洲篮球联合会主席)	
简介	姚明(Yao Ming),男,汉族,无党派人士,1980年9月12日出生于上海市徐汇区……	
属性表	名称	姚明
	性别	男
	民族	汉族
	……	……
其他有效信息	1980年9月12日傍晚,姚明出生于上海市第六人民医院,出生时体重5公斤,比普通新生儿重了几乎一倍……	

来,但是这是全部的网页数据,还需要进一步筛选。本文的抽取方法可以这样描述:如果该实体中的属性表中有三个及以上的人物预定义属性,那么该实体就是个人物实体。人物预定义属性是代表人物的属性,本文采用马进等[15]提出的12种属性作为预定义属性。最终抽取出85万条人物实体。为了验证抽取结果的正确性,本文随机抽取了属性数量为3的200个实体,结果全部为人物实体。同时,把“简介”项和“其他有效信息”项合并作为人物介绍。

2.2 扩充人物属性

为了进一步挖掘出潜在的人物属性,提高属性的丰富度。本文进行了如下操作:

(1)选取出现次数较多的属性。为了获得较多的数据量,本文对人物的属性表中属性出现的次数进行了排序,选取了占比前2%的属性,占比情况见表3。

(2)属性名称归一化。其中有的属性是同一种属性,例如“国家”和“国籍”本文把这类属性归为一类,并设立一个标准化的属性名称。

表3 各属性的占比、正确率、已经过滤后的正确率

Table 3 Proportion, accuracy and filtered accuracy of each attribute

属性名称	占比/%	正确率/%(100条)	添加标准化规则后的正确率/%
中文名	97.69	99	无规则
国籍	78.47	96	99
出生日期	70.30	94	100
出生地	57.01	92	98
民族	46.80	96	100
毕业院校	27.01	85	98
性别	26.64	99	100
外文名	17.89	91	97
逝世日期	12.56	90	100
身高	12.27	97	100
体重	9.70	96	100
运动项目	6.44	97	99
所属运动队	4.33	94	98
信仰	2.39	88	97
学历	2.59	88	100
场上位置	2.02	95	99

这样,本文在原有的12个属性的基础上又增添了8个属性。

2.3 属性值的标准化

属性表在建立后如果直接用来实现远程监督的话,会导致数据质量不高,因为属性表中的属性值会有一定的问题,这是由于粗文本标注的本身问题造成的,为此,本文为每个属性的属性值建立过滤规则,从属性表中过滤出正确的属性值来减轻这种标注问题。以下为发现的问题和标准化方案。

(1)错标:标注的属性值与属性无关,或标称了其他属性的属性值。例如将“身高”属性的属性值标成了“69 kg”。

对于这种问题,本文考虑固定属性值的格式进行抽取,例如限制体重的属性值为“数字+kg”,“数字+公斤”的格式,这样就会确保属性值完全正确。类似的还有日期、性别、民族、学校等。

(2)边界错误:除了属性值外又多了其他的部分。该问题普遍存在,例如“毕业院校”的属性值为“北京医科大学公共卫生专业”,而正确的属性值为“北京医科大学”。

对于这种问题,本文考虑把常见的边界删除,或者只保留有效部分,例如在“出生地”属性中将属性值种带有“出生于”,“生于”字样全部删除,得到最终属性值。

(3)多属性值问题:对于具有多个属性值的属性,属性表里标注的是粗文本,需要设立规则把它拆出来。例如“职业”属性值为“紫砂壶艺人、陶瓷艺人、工艺师”。

对于这种问题,需要把它转换成列表的形式。寻找分隔符然后分开即可,类似的属性还有毕业院校、国籍等。

(4)漏标问题:文中有该属性的属性值,但是属性表中没有。漏标问题分为两种,一种是单属性值的属性没有标,另一种是多属性值的属性没有标全。

对于这种问题,本文没有提出有效的改进手段,这是标注本身的问题。漏标问题存在于所有属性当中,本文随机抽取了100条数据,观察“性别”属性值在人物介绍中存在的情况下,没有标注的属性表占16%,可以发现并且占比很大。对于多属性值的属性则更高。

为了减轻这种影响,本文在评测中对召回

值R的计算公式做了一些修改,只考虑标注出来的属性值,对于没有标注的属性,无论预测成什么都不予考虑。这在第4节中会详细介绍。

(5)混合问题:这种问题为以上问题的混合形式,通常是在多属性值的情况下产生边界问题、漏标问题、属性值错误问题。例如“别名”的属性值为“字孔明、号卧龙”,还有职业和成就的大量漏标问题。

这种情况无法进行更改,所以本文把这种问题占比很大的属性删除,包括“职业”“主要成就”“别名”和“作品”。最终留下了16个属性。

本文对每一个属性,随机抽取了100条数据,记录属性值标注正确的比例,并记录在这些经过标准化后正确的比例,如表3所示。

2.4 属性值过滤和文本过滤

标注数据需要确保属性表中的属性值在人物介绍文本中,因此需要过滤掉不在文本中的属性删除。对于多属性值的属性,仅删除不在文本中的属性值。之后本文又对属性数量小于3的人物实体删除。最终获得了45万条数据。

为了获取高质量的人物介绍。本文在人物介绍中删除了没有出现属性值的句子,得到短文本的人物介绍。

标注的结果应以“属性表,人物介绍,短文本人物介绍”的形式给出。以“徐伯才”为例,如表4所示。标注数据的基本统计信息如表5所示。

为了进一步获取高质量的标注数据作为训练数据,本文对标注的数据质量进行打分,打分原则有两点(1)单位长度含有的属性越多分值越高。(2)含有稀有属性(占比低的属性)越多分值越高。考虑上述两点建立如下打分公式(1)。

$$S_{\text{person}}=\sum_{x\in I}P(x)+\frac{L(I)}{L(S)}\times\alpha, \tag{1}$$

其中I表示属性表(Infobox),S表示人物介绍(Summary),$P(x)$表示属性x出现次数的占比(Pencentage),$L(x)$表示x的长度,α为一个超参数。用来平衡两种打分原则的值的大小,在本实验中设为10。

表4 人物“徐伯才”的标注结果

Table 4 Annotated results of "Xu Bocai"

人物介绍	徐柏才,男,1953年11月生,汉族,湖北鄂州人,中共党员,1975年7月毕业于华中师范大学数学系,大学学历,教授,硕士研究生导师。曾任中南民族大学党委副书记。1998年10月—12月,在国家教育行政学院学习进修。 历任华中师范大学数学系学生政治辅导员、分团委书记,中南民族学院数学系党总支副书记、书记,中南民族学院党委组织部……	
短文本人物介绍	徐柏才,男,1953年11月生,汉族,湖北鄂州人,中共党员,1975年7月毕业于华中师范大学数学系,大学学历,教授,硕士研究生导师。曾任中南民族大学党委副书记。	
属性表	名称	徐柏才
	性别	男
	民族	汉族
	出生日期	1953年11月
	出生地	湖北鄂州
	学历	大学

表5 数据统计信息

Table 5 Statistic information

数据量	446 309
平均句长	7 315
短文本平均句长	829
平均属性数量	4.6

接着按照分值对数据项进行排序,由于分值过大的人物实体中的人物介绍与属性表的格式几乎相同(以表6为例)并不能作为有效的文本,所以弃掉了分值在前1 000的人物实体,选取排名在1 000到41 000的人物项作为训练数据。

表6 非有效标注数据

Table 6 Invalid annotated data

人物介绍	姓名:李治文性别:男 民族:汉族 出生日期:1953年11月 出生地:湖北鄂州 学历: 大学	
属性表	名称	李治文
	性别	男
	民族	汉族
	出生日期	1953年11月
	出生地	湖北鄂州
	学历	大学

3 基于阅读理解模型的属性抽取

3.1 模型介绍

本文提出了一种新的属性抽取任务模式,

传统的属性抽取采用序列标注模型进行抽取，由于属性重合的问题已经不能适应该新任务的需要。由于该任务本身与阅读理解模型相契合，输入都是一段文本，输出答案。为此本文将阅读理解模型应用于此数据。需要说明的是，本文仅针对该模式提出了在模式上比较契合的模型，对于该数据模式采用什么样的方法来抽取最有效仍需要大量的实验。

本文采用问答范式实现阅读理解模型，即把阅读理解表示成输入一段文本和想要问的问题，输出该问题的在文中的答案。如表7所示。

表7 人物“姚明”在阅读理解模型中的定义

Table 7 The definition of the character "Yao Ming" in the reading comprehension model

文本	姚明(Yao Ming)，男，汉族，无党派人士，1980年9月12日出生于上海市徐汇区……
问题	答案
该人物的姓名是什么?	姚明
该人物的性别是什么?	男
该人物的民族是什么?	汉族
该人物在哪天出生?	1980年9月12日

答案的搜索有两种方法，一种是基于跨度(Span)的方法，该方法需要对每一个字符添加两个分类器用于预测是否是实体的开始和结束的标志，同时需要再添加一个开始-结束匹配矩阵来预测两个分开的实体间是否需要连接组成新的更长的实体。另一种是基于序列标注的方法，该方法对每个字符添加一个多分类器用于预测该字符在实体中的位置。两种对比起来，前者能够解决嵌套实体的问题，后者会占用更少的计算空间。由于该任务的标注结果并没有嵌套实体的情况，因此本实验采用基于序列标注的阅读理解方法。

需要强调的是，这里的序列标注是为了表明使用token级别的阅读理解的方法，与之前提到的序列标注模型不同，两者虽然都采用token级别的标注，但序列标注模型对于同一个token只能给出一个标签，而基于序列标注的阅读理解的方法可以针对不同的理解目标对同一个token给出不同的标签，例如“美国加州”中“美”在前者只能被标为“B-国籍”或者“B-出生地”，而在后者两个标签全部可以标出。

本文分别使用BERT-MRC、BERT-CRF-MRC两个基准模型，CRF层通过考虑周围的特征来决定该字符的特征。在人物属性抽取中广泛使用。

3.2 基于BERT-MRC的属性抽取

BERT[11]是一种基于Trans-former的预训练模型，它通过采用Mask的方式实现了双向语言的表征。使用预训练+微调的方式训练，在各项任务中都得到了非常好的效果。

为了适应不同的任务，BERT的输入由一个开头占位字符[*CLS*]再加上多个句子组成，句子之间可以用[*SEP*]符号用来表示分隔的意义。在本实验中，本文沿用了Li等[14]的方法，将问题q和文本x用[*SEP*]符号拼接而成。组成如下字符串。

$$\{[CLS], q_1, q_2, \cdots, q_m, [SEP], x_1, x, \cdots, x_m\}$$

BERT的输入和输出是对应的，在本实验中只截取文本对应的输出，并在此基础上添加一个多分类器用于预测是否是实体。选取最大的类别作为该字符的类别。模型架构如图3所示。本实验采用交叉熵作为损失函数。

3.3 基于BERT-CRF-MRC的属性抽取

BERT-CRF-MRC与BERT-MRC的区别在于，前者代替了后者的线性分类器的做法，在经BERT输出后交给CRF层用来预测和训练标签。

CRF被广泛应用于序列标注模型中，它通过周围的标签来预测本身的标签。具体可描述为如下形式。给定输入序列$X=(x_1, x_2, x_3, \cdots, x_n)$，和与该序列等长的标签序列$Y=(y_1, y_2, y_3, \cdots, y_n,)$，定义该标注方案下的得分公式如式(2)所示。

$$S(x, y)=\sum_{i=0}^{n} A_{y_i, y_{i+1}}+\sum_{i=0}^{n} \boldsymbol{P}_{i, y_i}, \tag{2}$$

其中$A_{i,j}$表示从标签i转移到标签j的得分，$P_{i,j}$代表第i个输入状态标注为第j个标签的得分。在BERT-CRF中，得分矩阵M代表BERT层的输出结果。在训练时最大化条件概率中的正确标签序列的对数似然概率，如公式(3)所示。

$$M(y|x)=\frac{e^{S(x, y)}}{\sum_{\tilde{y} \in Y_x} e^{S(x, \tilde{y})}}, \tag{3}$$

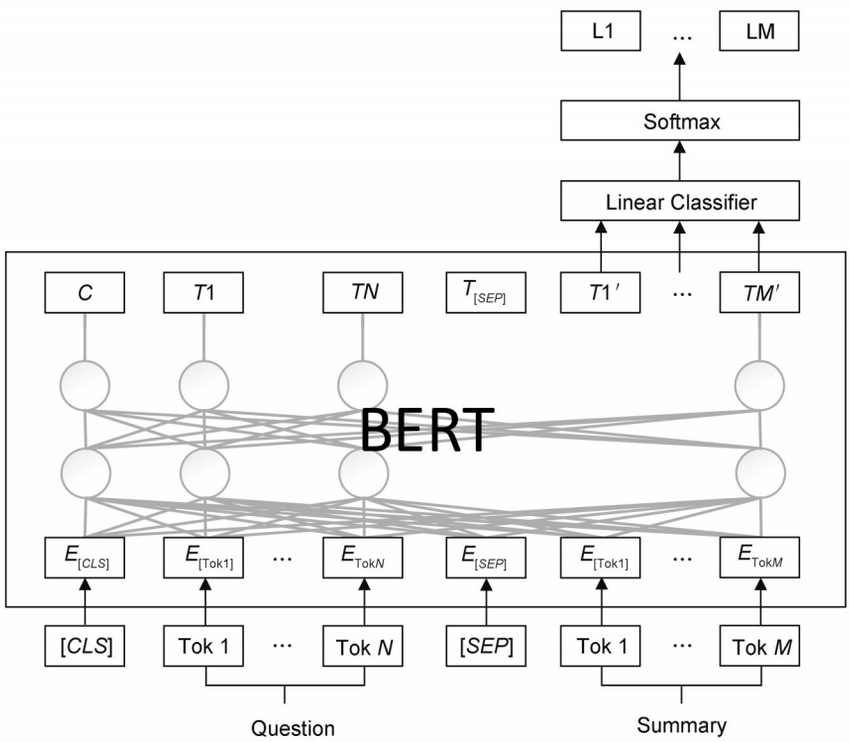

图3 BERT-MRC模型架构

Fig. 3 Framework of BERT-MRC

其中 Y_x 代表在当前序列下所有可能的标签序列,模型需要找到得分最高的序列。也就是找到满足公式(4)的 y。

$$\text{argmax}_{\tilde{y}\in Y_x} S(x,\tilde{y}) \tag{4}$$

3.4 输出属性值的归一化

由于本文的目的在于建立一个文档级别的语料库,因此在验证模型的好坏时应采用文档级别的验证方式,即对于每一个属性一对一地进行比较。但是本文提出的阅读理解模型会从文本中抽取出很多属于一种属性的属性值,以表8为例。

表8 模型抽出多个属性值的例子

Table 8 Example of extracting multiple attribute values from a model

文本	模型识别“出生地”的结果
谭云山,湖南人,1890年出生于湖南长沙	[“湖南”,“湖南长沙”]

在这个例子中两个都识别正确但后者包含的信息更多,后者应为最佳的正确答案。

基于这种情况,识别系统需要在模型输出后再添加一层属性值的归一化来获取模型输出的最佳属性值。为此,本文采用规则的方式,对每一个属性编写归一化规则。这种规则的种类有:(1)获取出现次数最多的属性值;(2)让子串相同的属性归为一类,然后返回出现次数最多的一类,该类的代表为最长的属性值;(3)先对属性值进行格式化(例如日期的年月日),然后选取格式化信息包含最多的属性;(4)多属性值的属性返回属性值的集合;(5)添加字符串限制,例如“民族”属性的末尾必须有“族”,性别只保留男或女。

4 实验

4.1 实验设置

本文采用BERT和BERT-CRF两种方法构建该任务的基线模型。将构造好的训练数据按8:1:1切分为训练集,测试集,开发集。根据Li等[14]的经验采用表9的参数设置,当在5次迭代中开发集的数据的 $F1$ 值不再提升时终止训练。

4.2 评价方法

本文在训练和评测时采用不同的评价方式。在训练时,模型依照上述模型提出的损失函数进行训练,采用短文本数据进行训练。在

表9 模型参数设置

Table 9 Parameter settings

参数	参数值
bert embedding	768
optmizer	adamw
learning rate	0.000 01
batch size	64
epoch	20

评测时,本文采用阅读理解的方式。

本实验对上述构建的长文本人物介绍和短文本人物介绍两种数据分别作出评价。

实验采用准确率(P)、召回率(R)和两者的调和平均值($F1$值)作为评价指标。对于每一个属性,评价指标的计算方式如式(5)—式(7)所示。

$$P=\frac{\text{模型识别成功的属性值数量}}{\text{模型识别的属性值数量}},\tag{5}$$

$$R=\frac{\text{模型识别成功的属性值数量}}{\text{测试集中属性值数量}},\tag{6}$$

$$F1=\frac{2PR}{P+R}。\tag{7}$$

为了减少漏标带来的误差,本文不考虑模型识别出但是本身没有进行标记的情况。即模型识别成功的属性值数量只有在属性表有该属性的前提下进行计数。

4.3 实验结果分析

本文共进行了四组实验,用训练好的BERT-MRC和BERT-CRF-MRC分别对长文本人物介绍和短文本人物介绍进行识别,实验结果如表10所示。

表10两种模型对长文本人物介绍和短文本人物介绍进行识别的对比实验结果,其中BERT-MRC表示基于BERT预训练语言模型的阅读理解模型,BERT-CRF-MRC表示在BERT-MRC的基础上添加了CRF(条件随机场)层进行解码

从表中可以发现如下信息:

(1)BERT-CRF-MRC要比BERT-MRC在$F1$值上平均高出三个百分点。

(2)短文本模型普遍取得较好的效果。$F1$值在90%以上的属性超过半数,这些属性的属性值构成的模式简单,并且触发词较为明显。

(3)长文本的$F1$值要比短文本低15个百分点。文本的增长会使模型识别出更多的属性值。这就会增加识别错误的概率。

对错标的数据进行观察,总结出如下原因:

表10 两种模型对长文本人物介绍和短文本任务介绍识别的对比实验结果

Table 10 Comparative experimental results of two models for recognizing long text character introductions and short text task introductions

属性	短文本			长文本		
	数据量	BERT-CRF识别率/%	BERT-CRF-MRC识别率/%	数据量	BERT-MRC识别率/%	BERT-CRF-MRC识别率/%
民族	638	99.52	99.52	329	78.03	78.46
性别	728	98.62	98.76	284	75.60	75.20
运动项目	1 042	95.29	98.16	321	95.29	95.61
出生日期	2 194	97.81	97.96	855	85.25	83.30
姓名	3 085	96	97.93	887	87.47	88.03
国籍	2 167	95.48	96.65	855	77.40	76.61
体重	728	93.92	96.4	203	78.35	80.44
逝世日期	501	94.3	93.53	296	76.58	83.04
场上位置	555	89.45	92.65	238	88.34	86.99
外文名	1 091	89.42	91.63	489	65.28	76.51
学历	612	88.33	90.23	201	50.12	50.89
所属运动队	821	65.57	89.7	315	75.85	75.24
出生地	1 831	75.57	89.55	802	62.11	63.05
毕业院校	919	81.61	85.98	711	59.90	63.05
信仰	537	71.05	79.02	232	54.58	58.64
身高	977	86.17	73.68	239	47.63	68.73
平均	1 151	88.63	91.95	453	72.36	75.15

(1)识别错误,即模型识别错误。

(2)标注本身错误。

(3)多属性值漏标。属性表没有标全的属性值被模型识别出来。例如学校。

(4)某些属性值有多种表达方式,例如“控球后卫”与“控卫”,“宾夕法尼亚州”与“美国宾夕法尼亚”。模型识别出的属性值和属性表中的属性值虽然字符不一样,但是两者描述的是同一属性值。

(5)有些属性值以最新的为准,而文本中会包含历史属性值信息,这些信息会被模型捕捉到。但并没有识别出最新信息,或者最新信息被归一化函数去掉了。例如学历。

本文随机抽取100个错误例子,计算不同类型错误的占比。如表11所示。从这两个表中可以发现如下信息:

本文随机抽取100个错误例子,计算不同类型错误的占比。如表11所示。

表11 不同类型标注错误的占比(100例)

Table 11 Proportion of different types of annotation errors (100 cases)

错误	典型属性	占比/%
识别错误	国籍、身高、出生地	46
标注错误	无	1
多属性值漏标	所属运动队、毕业院校	29
多种表达方式	场上位置,出生地,姓名	12
只识别旧属性值	学历,国籍,所属运动队	12

5 结论

本文提出基于阅读理解的标注方式以解决序列标注所带来的问题,并构造一份文档级别的人物属性抽取数据。该标注方式将数据构造成一段文本和一个属性表。本文用此方式构建了大规模长文本数据集和短文本数据集。并引入基于问答范式的BERT-MRC和BERT-CRF-MRC两个模型作为任务的基线。评测结果显示,模型在短文本上识别率较高,在长文本上识别率较低。

本文并没有考虑比较复杂的属性,例如职业,作品等。并且属性的归一化采用简单的规则方式。在未来工作中可以考虑采用限制属性值的长度,或者引入词典,将这些属性也加入进来。并且考虑提高属性的归一化质量,来提升系统准确率。

参考文献:

[1] 徐庆婷, 洪宇, 潘雨晨等. 属性抽取研究综述[J]. 软件学报, 2023, **34**: 690-711. DOI: 10.13328/j.cnki.jos.006709.
XU Q T, HONG Y, PAN Y C, *et al*. Survey on Aspect Term Extraction[J]. *J Softw*, 2023, **34**: 690-711. DOI: 10.13328/j.cnki.jos.006709.

[2] EMBAR V, KAN A, SISMAN B, *et al*. DiffXtract: Joint Discriminative Product Attribute-value Extraction[C]//2021 IEEE International Conference on Big Knowledge (ICBK). New York: IEEE, 2021: 271-280. DOI: 10.1109/ICKG52313.2021.00044.

[3] 李昊迪. 医学领域知识抽取方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2018.
LI H D. Research on Medical Domain Knowledge Extraction Methods[D]. Harbin: Harbin Institute of Technology, 2018.

[4] FAN Z F, WU Z, DAI X Y, *et al*. Target-oriented Opinion Words Extraction with Target-fused Neural Sequence Labeling[C]//Proceedings of the 2019 Conference of the North. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 2509-2519. DOI: 10.18653/v1/n19-1259.

[5] HU M Q, LIU B. Mining and Summarizing Customer Reviews[C]//Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2004: 168-177. DOI: 10.1145/1014052.1014073.

[6] 李红亮. 基于规则的百科人物属性抽取算法的研究[D]. 成都: 西南交通大学, 2013.
LI H L. Research on Character Attributes Extraction Based on Rules from Baidu Encyclopedia[D]. Chengdu: Southwest Jiaotong University, 2013.

[7] HOCHREITER S, SCHMIDHUBER J. Long Short-term Memory[J]. *Neural Comput*, 1997, **9**(8): 1735-1780. DOI: 10.1162/neco.1997.9.8.1735.

[8] JOHN L, ANDREW M, FERNANDO P. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//The 18th International Conference on Machine Learning. Williamstown, Massachusetts, USA: Morgan Kanfmann Publishers Inc. 2001: 282-289. DOI: 20.500.14332/6188.

[9] FAN Z F, WU Z, DAI X Y, *et al*. Target-oriented Opinion Words Extraction with Target-fused Neural Sequence Labeling[C]//Proceedings of the 2019 Conference of the North. Stroudsburg, PA, USA: Association for Computa-

tional Linguistics, 2019: 2509−2518. DOI: 10.18653/v1/n19-1259.

[10] DAI H L, SONG Y Q. Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 5268−5277. DOI: 10.18653/v1/p19-1520.

[11] VASWANI A, SHAZEER N M, PARMAR N, *et al*. Attention is All You Need[EB/OL]. (2017−06−12) [2025−04−21]. https://doi.org/10.48550/arXiv.1706.03762.

[12] DEVLIN J, CHANG M W, LEE K, *et al*. BERT: Pre-training of Deep BidirectionalTransformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2018. 4171−4186. DOI:10.18653/v1/N19-1423.

[13] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging[EB/OL]. (2015−08−09) [2025−04−21]. https://doi.org/10.48550/arXiv.1508.01991.

[14] LI X Y, FENG J R, MENG Y X, *et al*. A Unified MRC Framework for Named Entity Recognition[EB/OL]. (2019−10−25) [2025−04−21]. https://doi. org/10.48550/arXiv.1910.11476.

[15] 马进, 杨一帆, 陈文亮. 基于远程监督的人物属性抽取研究[J]. 中文信息学报, 2020, **34**(6): 64−72. DOI: 10.3969/j.issn.1003-0077.2020.06.009.
MA J, YANG Y F, CHEN W L. Distant Supervision for Person Attribute Recognition[J]. *J Chin Inf Process*, 2020, **34**(6): 64−72. DOI: 10.3969/j.issn.1003-0077.2020.06.009.

[16] 张巧, 熊锦华, 程学旗. 基于弱监督学习的主页人物属性抽取方法[J]. 山西大学学报(自然科学版), 2015, **38**(1): 8−15. DOI: 10.13451/j. cnki. shanxi. univ (nat.sci.).2015.01.002.
ZHANG Q, XIONG J H, CHENG X Q. Person Attributes Extraction Based on a Weakly Supervised Learning Method[J]. *J Shanxi Univ Nat Sci Ed*, 2015, **38**(1): 8−15. DOI: 10.13451/j.cnki.shanxi.univ(nat.sci.).2015.01.002.

[17] ANGELI G, TIBSHIRANI J, WU J, *et al*. Combining Distant and Partial Supervision for Relation Extraction[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1556−1567. DOI: 10.3115/v1/d14-1164.

[18] 苏丰龙, 谢庆华, 邱继远, 等. 基于深度学习的领域实体属性词聚类抽取研究[J]. 微型机与应用, 2016, **35**(1): 53−55. DOI: 10.19358/j.issn.1674-7720.2016.01.017.
SU F L, XIE Q H, QIU J Y, *et al*. Study on Word Clusting for Attribute Extraction Based on Deep Learning[J]. *Microcomput Appl*, 2016, **35**(1): 53−55. DOI: 10.19358/j.issn.1674-7720.2016.01.017.

[19] 向晓雯. 基于条件随机场的中文命名实体识别[D]. 厦门: 厦门大学, 2006.
XIANG X W. Chinese Named Entity Recognition Based on Conditional Random Fields[D]. Xiamen: Xiamen University, 2006.

[20] KATIYAR A, CARDIE C. Investigating LSTMS for Joint Extraction of Opinion Entities and Relations[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 919−929. DOI: 10.18653/v1/p16-1087.

[21] CHO K, VAN MERRIENBOER B, BAHDANAU D, *et al*. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[C]//Proceedings of SSST−8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 103−111. DOI: 10.3115/v1/w14-4012.

[22] RADFORD A, NARASIMHAN K, SALIMANS T, *et al*. Improving Language Understanding by Generative Pre-Training [J]. *Open Access Library Journal*, 2021, **8**: 7.

[23] PETERS M E, NEUMANN M, IYYER M, *et al*. Deep Contextualized Word Representations[EB/OL]. (2018−02−14) [2025−04−21]. https://doi. org/10.48550/arXiv.1802.05365.