基于远程监督的人物属性抽取研究*

马进[1]，杨一帆[1]，陈文亮[1]

（1.苏州大学 计算机科学与技术学院，江苏省 苏州市 215006）

摘要：属性抽取的主要目标是从非结构化文本中获取实体的属性值。为了从文本中抽取出人物属性，通常需要大量的标注数据，然而这些数据资源却十分稀少。为了解决这个问题，本文从百科类网页的表格数据出发，构建了人物属性表，然后采用远程监督的方法得到大规模、多类别的人物属性标注语料，从而免去了人工标注的繁琐流程。针对新构建的数据集，分别使用条件随机场（CRF）和双向长短期记忆-条件随机场（BiLSTM-CRF）构建了属性抽取的两个基线模型。实验结果表明 BiLSTM-CRF 取得比 CRF 更好的性能，其中 BiLSTM-CRF 的平均 F1 值为 83.39%。

关键词：属性抽取；标注数据；远程监督

中图分类号：TP391　　**文献标识码：**A

Distant Supervision for Person Attribute Recognition

Ma Jin[1], Yang Yifan[1], CHEN Wenliang[1]

(1. School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, 215006, China)

Abstract: Attribute recognition aims at obtaining attribute values of entities from unstructured text. In order to extract person attributes from text, a large amount of annotated data is usually needed. However, there is a lack of a massive amount of such data so far. To address this issue, we use infobox of encyclopedia web pages to construct the tuples of person attributes, and then use distant supervision method to obtain large-scale and multi-category annotated datasets for person attributes, thus avoiding the tedious process of manual annotation. Additionally, we present two kinds of models based on CRF and BiLSTM-CRF for person attribute recognition as the baseline systems. The experimental results show that BiLSTM-CRF performs better than CRF on this newly built dataset.

Key words: Attribute Recognition; Annotated data; Distant supervision

0 引言

随着互联网、大数据等计算机技术的发展，全球数据量呈现爆炸式的增长。据统计，截止到 2014 年，全球互联网上被保存下的数据量达到 50EB，并且每两年将翻一倍。但是，互联网上掺杂着大量混乱的非结构化数据，如何利用这些庞大的数据，从中获取有用的知识，是一项值得我们思考的工作。非结构化数据（Unstructured data）[1]是指不含有预先定义的数据模型（Data model）[2]，即没有被组织成预先定义好形式的信息。百科类的网页中包含着大量非结构化的文本数据，具有非常丰富的实体属性信息。

属性抽取的定义是：给定一个实体及其属性列表，从一个非结构化的文本中抽取出该实体的各个属性值。例如有一段关于实体“姚明”的部分词条简介，如表 1 中最上方所示，从这段文本可以获取到如表 1 的下半部分所示的人物属性值。

属性抽取的应用非常广泛。一方面，属性抽取是构建知识图谱的关键子任务，可以用于知识图谱的补全和纠错。另一方面，可以用来挖掘人们感兴趣的相关属性，例如人的出生地、组织的成立时间、疾病的症状等等。

现有的属性抽取方法主要有基于规则的方法[3]和传统机器学习的方法[4]。基于规则的方法首先要手工构造模式，然后利用这些模式去匹配大量文本，匹配到的结果即该人物的属性值。基于传统机器学习的方法一般使用有监督的学习策略，但是该方法需要大规模的标注语料。近几年来，深度学习的方法在自然语言处理的各个任务上已经证明有效，如机器翻译[5]、情感分析[6]、关系抽取[7]等。循环神经网络（Recurrent Neural Network，RNN）[8]及长短期记忆网络（Long Short-Term Memory，LSTM）[9]和门控网络（Gated Recurrent Unit，GRU）[10]在序列标注建模上有突出表现。但是，深度学习的方法同样需要大量标注语料来训练模型。

收稿收稿日期：0000-00-00；**定稿日期：**0000-00-00

基金项目：国家自然科学基金（61525205，61876115）

作者简介：马进（1995），男，硕士研究生，主要研究方向自然语言处理；杨一帆（1995），男，硕士研究生，主要研究方向自然语言处理；陈文亮（1977），男，教授，主要研究方向自然语言处理。

目前，属性抽取任务在中文上鲜有公开的大规模标注数据集。研究者一般研究特定领域或者面向工程应用，注重任务本身，从无到有构建自己的数据进行实验研究。并且随着深度学习技术的发展，属性抽取任务对数据量的要求越来越高。为了解决这些问题，一种直接的方法是人工标注数据构建完整的数据集，但这需要花费大量的时间以及人力成本，同时标注员的标注水平也会很大程度地影响标注语料的质量。出于这样的原因，本文使用远程监督的方法，自动匹配生成标注数据，构建了一个大规模人物属性抽取数据集，涵盖 12 种类别的属性和总计 147 万条的标注语料。

表 1 实体“姚明”部分简介及部分可获取的属性

介绍	姚明（Yao Ming），男，汉族，无党派人士，1980年9月1日出生于上海市徐汇区，祖籍江苏省苏州市吴江区震泽镇，前中国职业篮球运动员，司职中锋…			
性别	民族	出生时间	出生地点	祖籍
男	汉族	1980年9月12日	上海市徐汇区	江苏省苏州市吴江区震泽镇

本文将属性抽取任务看作序列标注问题，图 1 是我们采用的任务框架图。首先从百科类网页上获取大量非结构化文本与表格数据，并对该数据进行一系列复杂的抽取得到人物属性表。接着使用远程监督的方法并结合现有人物属性表在文本中自动匹配生成标注数据。最后，使用条件随机场（CRF）[11]和双向长短期记忆-条件随机场（BiLSTM-CRF）[9]进行建模，获得两种人物属性抽取系统。

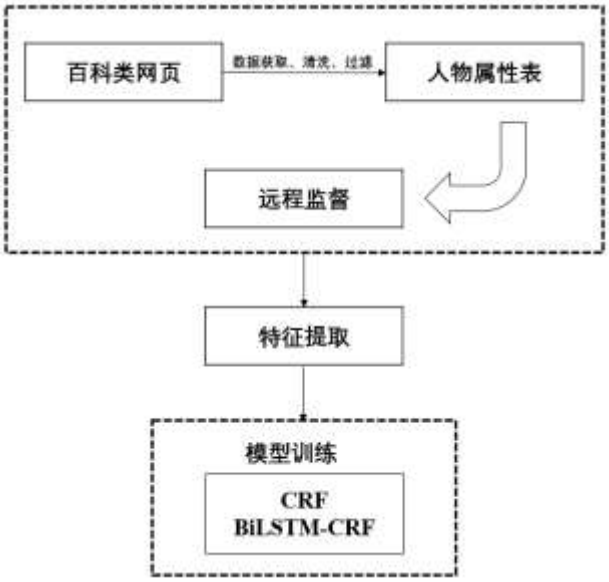

图 1　本文采用的属性抽取任务框架

实验结果表明，相较于传统用于解决序列标注任务的条件随机场，深度模型在各个属性上都能取得更好的效果。综上，本文的贡献如下：

1) 利用远程监督的方法，自动构建了大规模、多类别的人物属性标注语料库。获取的中文人物属性标注语料库（CPAR）已向业界公开免费使用（https://github.com/SUDA-HLT/CPAR）。
2) 构建了两种基线模型。用传统 CRF 方法构建人物属性抽取任务传统方法的基线模型，更进一步使用深度学习与传统方法相结合的 BiLSTM-CRF 方法建模，作为深度学习方法的基线模型。

1 相关工作

属性抽取任务由来已久，然而目前并没有一个标准的公开数据集供研究者进行实验研究。工作者一般都是立足于任务本身，从无到有构建数据集。朱臻等人[12]从七个网站上获得总计 2400 条、涵盖 4 种属性关系的藏语语料进行藏语人物属性抽取。张巧等人[13]从美国十所大学页面上获取 810 张、包含 9 种属性的导师主页构建了英文语料库用于主页人物属性抽取。张丙奇等人[14]从企业网页上获取企业的基本信息，构建了涵盖 8 个属性的小规模企业信息抽取技术中文数据集。Angeli[15]基于众包构建了一些干净、高质量的标注数据。这些数据集都存在不足的地方，首先，这些数据不够统一，很难重复利用。其次，构建数据集的过程不仅繁琐耗时，而且往往会耗费大量人力物力成本。另外，这些数据并没有开源，研究者无法用这些数据进行更深一步的对比研究。

目前关于属性抽取的任务研究主要集中在算法方面，主要有基于规则和基于机器学习的方法。Hearst [16]为了寻找下义词的关系，构造了一系列的模式。该方法虽然可以完成属性抽取的任务且能够保证准确率，但是模式难以构造，维护非常困难，而且构造的模式大多是与领域相关，移植难度大。由于这些缺陷，有人提出自举的方法[17]，即半监督的方法。该方法在有一些实例和模式的情况下，迭代地生成新的实例和模式。Brin 等人[18]1998 年用这种方法提出 DIRPE 系统用于抽取作家与出版书籍之间关系。

基于机器学习的方法分为有监督和无监督的方法。Kambhatla[19]使用了词法和句法两类特征，这两种特征是分别从句法解析树和语法依赖树当中提取的。GuoDong 等人[20]在 Kambhatla 的基础上，深入细化使用单词和句法信息，加上词组、WordNet 和名字列表等信息。Lodhi 等人[21]提出字符串核函数（String Kernels）的概念，这种思想也很快被应用

到有监督的属性抽取任务上。Hasegawa 等人[22]首先提出无监督的方法，Chen 等人[23]针对 Hasegawa 的方法中存在的问题，提出了新的方法。该方法不需要手动标注关系实例及定义聚类个数，可以避免提取各个类别的标签，从而不会造成对类别信息的偏移。Huang[24]提出一种不同的方法，该方法使用基于神经网络的独立图作为输入并伴以两种注意力机制，能够更好捕捉到指示性信息。Rajani [25]也尝试组合来自多个系统的结果来确定属性类型。

本文基于百科人物介绍和表格数据（InfoBox），将属性抽取任务作为序列标注问题进行处理。早期阶段，序列标注任务的相关工作大多采用 CRF 模型进行。近年来，基于深度学习的方法在一些序列标注任务上取得了很好的效果。基于 CRF 模型的序列标注模型通常需要人工构建复杂的特征模板。近三年来，随着深度学习的发展，更多的研究人员采用深度学习的序列标注方法进行建模，如 Irsoy 等[8]基于 RNN 模型，Katiyar 等[9]用到 LSTM 模型。

2 人物属性标注语料库构建

2.1 数据获取

本文实验数据来自百科，在数据获取阶段，一方面要从网页的介绍文本中获取实体的词条简介，另一方面还需要从网页的表格信息中抽取出实体三元组。为了获取完整的实体信息，我们首先浏览一部分网页，以便定位 InfoBox 在 HTML 页面中位置，其次也需准确定位 InfoBox 中的每一条目的位置。在发现位置规律后，采用基于规则的方法，将 InfoBox 信息框中的结构化信息和实体对应的介绍文本从 HTML 网页中全部抽取出来。

以篮球运动员“姚明”为例，该实体三元组抽取的部分结果如表 2 所示，实体词条部分简介如表 3 所示。对于每个属性或者关系，都能够以**<姚明，星座，处女座>**这样的形式进行表述。接着，对得到的数据进行去重和属性归一化处理。最终，得到约 893 万个实体及其对应的词条简介，3828 万条三元组。

表 2 实体“姚明”三元组抽取结果

词条 基本信息	中文名	姚明
	外文名	Yao Ming
	国籍	中国
	出生地	上海市徐汇区
	……	
	生肖	猴
	位置	中锋
	妻子	叶莉
	星座	处女座
词条标签	运动员，篮球，体育人物	

表 3 实体“姚明”词条简介抽取结果

词条 简介	姚明(Yao Ming)，1980年9月12日出生于上海市徐汇区，祖籍江苏省苏州市吴江区震泽镇，前中国职业篮球运动员，司职中锋，现任中国篮球协会主席、中职联公司董事长兼总经理。 1998年四月，姚明入选王非执教的国家队，开始篮球生涯。2001年夺得CBA常规赛MVP，2002年夺得CBA总冠军以及总决赛MVP，分别三次当选CBA篮板王以及盖帽王，2次当选CBA扣篮王。在2002年NBA选秀中，他以状元秀身份被NBA的休斯敦火箭队选中，2003-09年连续6个赛季（生涯共8次）入选NBA全明星阵容，2次入选NBA最佳阵容二阵，3次入选NBA最佳阵容三阵。2009年，姚明收购上海男篮…

百科网页中抽取到的实体种类复杂，不仅包含人物实体，还有如机构、景点等其它类型的实体。为了筛选出人物实体，本文利用词条标签及常见人物属性，综合判别当前对象是否为人物实体。通过统计所有实体 InfoBox 中的属性及标签出现的频数，人工筛选出人物实体的常见属性和标签，用它们作为筛选标准。

表 4 人物实体的常见属性及标签

属性	频数	词条标签	频数
国籍	735,080	人物	781,199
职业	466,398	政治人物	167,416
民族	421,832	学者	102,575
性别	314,391	体育人物	97,445
毕业院校	228,019	娱乐人物	71,672

表 4 列举了本文选取的用来判断人物实体的几种常见属性及标签。在确定这些属性及标签后，我们采用如下两个规则筛选过滤出人物实体：

1) 若实体三元组中包含“国籍”、“职业”、“民族”、“性别”、“毕业院校”中任意一个属性，即将其归为人物实体。
2) 若实体词条标签里包含“人物”、“政治人物”、“学者”、“体育人物”、“娱乐人物”中任意一个标签，即将其归为人物实体。

从表中能看出，关联到人物实体的属性和标签总体数量可观。最终，我们得到约 111 万条人物实体，并对结果随机采样 200 条进行人工评估，其中 100%都是人物实体。

2.2 远程监督

本文采取远程监督的方法生成标注数据，该方法减少了有监督学习中人工标注数据的成本。远程监督方法是基于这样的假设：如果一个句子含有一个字串等于某个属性涉及的属性值，那么这个字串作为该属性的一个实例。

表 5 实体“姚明”部分词条简介的标注结果

词条简介	姚明（Yao Ming），男，汉族，无党派人士，1980年9月1日出生于上海市徐汇区，祖籍江苏省苏州市吴江区震泽镇，前中国职业篮球运动员，司职中锋…
标注结果	姚明（Yao Ming），**[男/性别]**，**[汉族/民族]**，无党派人士，**[1980 年 9 月 12 日/出生日期]**出生于**[上海市徐汇区/出生地]**，祖籍江苏省苏州市吴江区震泽镇，前中国职业篮球运动员，司职中锋…

在本实验中，远程监督是基于表 2 的三元组以及表 3 中的人物词条简介。与传统的远程监督方法不同，本文固定了头实体，只需在人物词条简介中寻找尾实体并标记，并不需要头实体和尾实体同时出现在句子中。这种方法之所以可行，是因为人物词条简介描述的主体对象就是头实体，描述的内容都与其相关。以实体“姚明”为例，得到的标注结果如表 5 所示。

2.3 数据规模

使用远程监督的方法生成标注数据，不仅减少人工标注数据的人力成本，还能获得大规模的有标注数据。表 6 列出了本数据集中的 12 种属性、对应远程监督的标注数据量及相应的采样准确率。

从表中可以直观看出，远程监督方法得到的有标注数据类别多样且规模庞大，同时，正确率也比较高。主要原因是该方法实际上进行了双层过滤。首先，人物实体 InfoBox 中存在的属性本就有一定准确性。其次，在满足前一个条件的情况下，相应的属性值还需要出现在人物词条简介中。经过这两个阶段筛选过滤，得到的标注数据正确率很高。

表 6 各属性及其标注数据量统计信息

属性	标注数据量	采样200条正确率
出生日期	413,319	100%
国籍	331,291	98%
出生地	321,640	99%
职业	293,345	99%
性别	162,278	99%
毕业院校	157,424	100%
民族	152,843	98%
逝世日期	64,757	100%
别名	37,579	99%
学位	33,110	97%
身高	2,061	100%
体重	1,247	100%

3 基于序列标注模型的属性抽取

3.1 模型特点

本文使用当前语料训练得到的字向量作为输入表达，这种预训练字向量非常贴合本文的属性抽取研究工作。一方面这种方法能够减少特征选择的过程，降低属性抽取任务的时间复杂度。另一方面，字向量是一种抽象的表达方式，不但能表达词义关系、词法信息，而且还能包含语义关系甚至是句法关系。

本文分别使用条件随机场（CRF）和双向长短期记忆-条件随机场（BiLSTM-CRF）构建属性抽取的两个基准模型。CRF 模型更多考虑的是句子的局部特征，通过特征模板去扫描整个句子。BiLSTM-CRF 模型是一种特殊的递归神经网络，可以有效解决训练过程中梯度消失和梯度爆炸的问题，并且能够处理输入信息间的序列信息。它的输出不但与当前输入有关，还会考虑到上下文的信息。

3.2 基于 CRF 的属性抽取

CRF 是一种判别式的概率无向图模型，其中线性链条件随机场（Linear chain Conditional Random Fields）被广泛用在序列标注任务中。它可以将序列标注任务转化成如下形式：给定输入序列如：$X=(X_1,X_2,\dots,X_n)$，目标是预测与该序列等长的标签序列$Y=(Y_1,Y_2,\dots,Y_n)$，标签序列中的每个位置和输入序列的每个位置一一对应。然后，由式（1）计算条件概率$P(\boldsymbol{y}|\boldsymbol{x})$：

$$P(\boldsymbol{y}|\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x})}\exp\left(\sum_{i,k}\lambda_k f_k(y_{i-1},y_i,\boldsymbol{x})+\sum_{i,k}\mu_k g_k(y_i,\boldsymbol{x})\right) \quad (1)$$

其中，$Z(\boldsymbol{x})$ 是归一化因子，f_k 和 g_k 分别是转移特征函数和状态特征函数，它们的输出值为 0 或者 1。对于f_k，当 y_{i-1}，y_i，$\boldsymbol{x}$ 满足转移特征函数的具体数值时输出为 1，否则为 0；g_k 取值与 f_k 类似。λ_k 和 μ_k 是对应特征函数的权值。在训练过程中，由输入序列和标签序列构成的每一组实例通过最大化式（1）的对数似然概率来训练模型的各个变量。测试时，给定测试数据中的一组输入序列实例 $\boldsymbol{x}'$，选取满足式（2）的输出序列 $\boldsymbol{y}^*$ 作为最佳预测标签序列：

$$\boldsymbol{y}^*=\operatorname{argmax} P(\boldsymbol{y}'|\boldsymbol{x}') \quad (2)$$

在本实验中，设计的 CRF 模型的部分特征模板如表 7 所示。其中，x 代表与当前位置的偏移量（0 表示当前位置），char 代表序列。F1 表示当前字、当前字的前后各三个单字作为特征，F2 表示当前字的前两个字、当前字和前一个字、当前字和后一个字、当前字的后两个字组成的特征，F3 表示当前字的前一个字和当前字的后一个字组成的特征。以“姚明出生于上海市”、当前位置词“生”为例，生成的特征如表中的 f1、f2 和 f3 所示。

表 7　CRF 模型采用的部分特征模板及特征举例

F1	char[x], x∈{-3,-2,-1,0,1,2,3}
F2	char[x]%char[x+1], x∈{-2,-1,0,1}
F3	char[x-1]%char[x+1], x∈{0}
f1	姚；明；出；生；于；上；海
f2	明%出；出%生；生%于；于%上
f3	出%于

3.3 基于 BiLSTM-CRF 的属性抽取

图 2 是本实验采用的 BiLSTM-CRF 框架图。第一层是数据表示层。其作用是将输入的字序列映射到向量级别的输入表示。需要说明的是，本文使用的是专门针对百度百科文本训练好的字向量，不仅能够有效解决未登录词问题，还能提高实验性能。

第二层是 BiLSTM 层。其作用是将第一层得到的输入表示转化为隐层表示输出。长短期记忆网络（LSTM）是一种特殊的循环神经网络（RNN）模型，能够学习长期的依赖关系，在处理序列数据时被广泛使用。

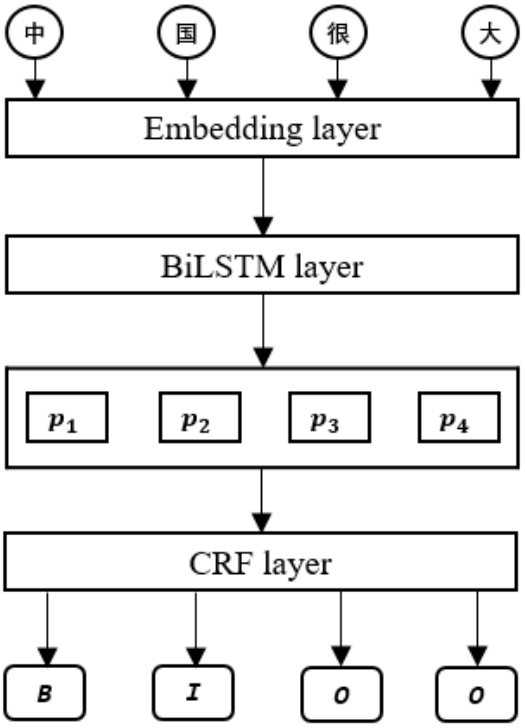

图 2　BiLSTM-CRF 框架图

模型的输入有三个：当前时刻的输入值X_t、上一时刻隐藏层输出值H_{t-1}、以及上一时刻的单元状态C_{t-1}。模型内部有三个控制开关，一个称为输入门I_t，在这个阶段决定保留多少X_t到C_t；一个称为遗忘门F_t，在这个阶段主要是对上一个节点传进来的输入进行选择性的遗忘。其主要结构如式（3）：

$$\begin{aligned} I_t &= \sigma(X_t \cdot W_{xi} + H_{t-1} \cdot W_{hi} + b_i) \\ F_t &= \sigma\left(X_t \cdot W_{xf} + H_{t-1} \cdot W_{hf} + b_f\right) \\ O_t &= \sigma(X_t \cdot W_{xo} + H_{t-1} \cdot W_{ho} + b_o) \\ \widetilde{C_t} &= \tanh(X_t \cdot W_{xc} + H_{t-1} \cdot W_{hc} + b_c) \\ C_t &= F_t \odot C_{t-1} + I_t \odot \widetilde{C_t} \\ H_t &= O_t \odot \tanh(C_t) \end{aligned} \quad (3)$$

隐层的最终表示 $H_t = \left[\overrightarrow{F_t}, \overleftarrow{B_t}\right]$，其中$\overrightarrow{F_t}$由 LSTM 对于一段输入时间序列从左至右进行计算得到，$\overleftarrow{B_t}$是从右至左进行计算得到。

第三层是 CRF 层，用于解码。假设序列标注的标签个数为 m，对于输入序列 $X = (X_1, X_2, \ldots, X_n)$，经过这三层的计算后可以得到维数为 $n \times m$ 的分值矩阵 P，矩阵中的某一个元素 $P_{i,j}$ 代表第 i 个输入状态标注为第 j 个标签的得分。对于一组预测标签序列 $Y = (Y_1, Y_2, \ldots, Y_n)$，定义它的得分如式（4）：

$$Score(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \quad (4)$$

其中，A 是转移得分矩阵，$A_{i,j}$ 代表从标签 i 转移到标签 j 的得分。y_0和y_n分别是标签序列中的起始和结束标签，需要加入到标签集合中。因此 A 是 $m+2$ 阶方阵。由此，我们得到基于所有可能的标签集合 $\boldsymbol{Y_x}$ 下的条件概率 $P(\boldsymbol{y}|\boldsymbol{x})$ 如式（5）：

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{e^{Score(\boldsymbol{x}, \boldsymbol{y})}}{\sum_{\tilde{\boldsymbol{y}} \in \boldsymbol{Y_x}} e^{Score(\boldsymbol{x}, \tilde{\boldsymbol{y}})}} \quad (5)$$

在训练过程中，最大化如式（5）中正确标签序列的对数似然概率。测试时，选取满足式（6）的结果 $\boldsymbol{y}^*$ 作为最佳预测标签序列：

$$\boldsymbol{y}^* = \underset{\tilde{\boldsymbol{y}} \in \boldsymbol{Y_x}}{\operatorname{argmax}} Score(\boldsymbol{x}, \tilde{\boldsymbol{y}}) \quad (6)$$

4 实验

4.1 交叉验证

交叉验证法（Cross Validation）的思想是先将数据集 D 划分为 k 个大小相似的不相交子集，满足 $D = D_1 \cup D_2 \cup \ldots \cup D_k$，$D_i \cap D_j = \emptyset (i \neq j)$。每个子集 D_i 都尽可能保持数据分布的一致性，都从 D 中通过分层采样得到。接着，每次用 k - 1 个子集的并集作为训练集，剩下的那个子集作为测试集。由此可以获得 k 组训练-测试集，从而可以进行 k 次训练和测试，最终对这 k 次结果取平均得到最终结果，本文采用五折交叉验证。

4.2 实验设置

本次实验，使用两种方法构建属性抽取的基线模型：CRF 和 BiLSTM-CRF。对于 CRF 模型，本文使用第三节设计的特征模板，在训练中，使用五折

交叉验证的方法，当满足设置的收敛条件时迭代终止。测试阶段，使用训练好的模型去测试集上进行五折验证。对于 BiLSTM-CRF 模型，我们按照表 8 所示的参数进行实验。使用五折交叉验证方法，将训练集中切分出 1/8 作为验证集，进行五十轮迭代，用在验证集上效果最好的模型去测试集上计算结果。

表 8　BiLSTM-CRF 模型超参数设定

Parameter	Value
emb size	300
max length	250
optimizer	Adam
learning rate	0.001
batch size	64
epoch	50
dropout	0.5

表 9　实验结果

属性	标注数据量	CRF			BiLSTM-CRF		
		P	***R***	***F1***	***P***	***R***	***F1***
性别	162,278	99.25	99.79	99.52	99.62	99.71	99.66
国籍	331,291	96.56	97.29	96.92	97.66	98.16	97.91
身高	2,061	93.24	90.14	91.66	93.15	91.36	92.25
体重	1,247	94.17	90.29	92.19	93.15	90.92	92.03
出生日期	413,319	89.94	89.87	89.90	91.10	90.95	91.02
民族	152,843	86.36	85.67	86.01	87.10	86.67	86.88
逝世日期	64,757	84.57	83.35	83.95	86.87	86.13	86.50
毕业院校	157,424	76.44	79.19	77.79	84.04	82.85	83.44
出生地	321,640	76.37	68.25	72.08	80.97	78.40	79.67
学位	33,110	75.10	70.86	72.92	78.10	77.53	77.81
别名	37,579	45.20	32.51	37.82	62.82	53.21	57.62
职业	293,345	52.58	35.25	42.20	58.19	53.66	55.83
平均				78.58			83.39

4.3 评价方法

实验结果采用识别准确率（*P*）、召回率（*R*）、和二者的调和平均 *F1* 值作为评价标准。对于每一属性，*P* 指正确识别的属性占总计识别出的属性的百分比，*R* 指正确识别的属性占测试集中所有属性的百分比，*F1* 是 *P* 和 *R* 的调和均值，可以综合考量模型的性能。*P*（准确率）、*R*（召回率）、*F1* 的计算方式如式（7）、（8）、（9）所示：

$$P = \frac{|A \cap G|}{|A|} \quad (7)$$

$$R = \frac{|A \cap G|}{|G|} \quad (8)$$

$$F1 = \frac{2PR}{P+R} \quad (9)$$

其中，$|A|$ 代表识别出的属性值总数，$|G|$ 代表标准集的属性值总数，$|A \cap G|$ 代表识别出的属性与标准集完全匹配的属性值总数。

4.4 实验结果与分析

本文共进行 2 组实验。第一组实验使用传统机器学习方法 CRF，作为传统方法的基线模型。另一组实验使用 BiLSTM-CRF，作为深度学习与传统方法相结合的基线模型。实验结果如表 9 所示。从表中可以看出：

1) 在“性别”、“国籍”这两个属性上效果最佳。一方面因为标注数据规模大，另一方面这些属性值构成简单，模型效果较好。
2) 在面对简单的属性抽取任务时，如“身高”、“体重”等，虽然标注数据不太多，两个模型也能有比较好的效果，这主要得益于该属性值特征比较强，模型容易学习。
3) 在面对复杂的属性抽取任务时，如“别名”、“职业”等，不管是 CRF 还是 BiLSTM-CRF 取得的效果都不是太理想。事实上，可以考虑加上 ELMo[26]或者 BERT[27]，不但能够考虑到句法和语义信息，还能对多义词进行建模。
4) BiLSTM-CRF 模型整体性能优于 CRF 模型，表明深度学习模型更适合本任务。

通过分析个例，会发现如下问题，例如对于“出生地”这个属性，有条标注数据“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]，祖籍江苏省苏州市吴江区震泽镇。”，其对应的无标注数据在经过本文的模型后，会得到“出生地”的属性值为“上海市”。虽然从直观上来

看这样的识别结果没有任何问题，但是在计算 *P*、*R*、*F1* 值时，这样的识别结果会被认作是错误的识别结果，因为它与标注数据并不完全一致。

以上这种部分识别的情况在各个属性上都有出现，在复杂属性上尤为明显。此外，针对识别效果不太理想的“别名”和“职业”这两个属性，分析如下。

对于“别名”这一属性，由于上下文表示形式多变以及实体别名的特殊性等原因，导致识别结果并不理想。例如，“慧宽尊者，唐代高僧。慧宽，一作惠宽”，这里面“一作”和“也叫作”意思相同，但是由于这种表达方式十分稀少，前后特征很难被模型捕捉，导致候选实体无法被识别出。还有一些识别错误的是把实体本名识别成实体的别名，以及识别结果正确，但是并未标注出，这些情况在评估时系统都会将其当成错误的识别结果。

对于“职业”这一属性，由于属性本身特点会导致识别的属性值当中会包含部分修饰词，例如“某某公司总经理”、“某某市副市长”、“自治区总工会副主席、党组成员兼自治区文联副主席、职工文联主席”等当中都出现不同程度的修饰词，这些修饰词不会经常出现，导致模型难以捕捉到上下文的语义特征，最终系统无法准确识别候选实体，出现识别不完整或者多识别的现象。另外，由于中文表达的复杂多样性，模型很难把握对应的语义信息，有时会出现识别不出相应的属性值的情况。例如一段描述“王秀勇原籍山东菏泽，双手残疾，以卖艺为生”，实体“王秀”的职业是“卖艺”，然而模型却识别不出正确结果。

5 总结

属性抽取任务需要大规模有标注数据，通常情况下需要人工标注来获得。为了解决这个问题，本文使用远程监督的方法，构造了大规模、多类别的人物属性标注语料，从而免去人工标注的繁琐流程。针对新构建的数据集，分别使用条件随机场和双向长短期记忆-条件随机场构建了属性抽取的两个基线模型。本文构建的数据集已开源，可以通过 https://github.com/SUDA-HLT/CPAR 获得，供相关研究人员使用。

属性抽取任务用途十分广泛，是构建知识图谱的关键子任务。本文没有使用复杂的模型去提升属性抽取的效果，在模型的设计上还有很大的提升空间。在未来工作中，会考虑加入外部资源或者使用适用于远程监督的模型[28]来改进系统性能。并且还会考虑做跨领域的属性抽取任务。

参考文献

[1] Rao R. From unstructured data to actionable intelligence[J]. It Professional, 2003, 5(6):29-35.

[2] Banerjee J, Chou H T, Garza J F, et al. Data model issues for object-oriented applications[J]. Acm Transactions on Information Systems, 1987, 5(1):3-26.

[3] 基于规则的百科人物属性抽取算法的研究[D]. 西南交通大学, 2013.

[4] 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.

[5] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[6] Wang J., Yu L., Lai K., et al. Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model[C]//In Proceedings of the Meeting of the Association for Computational Linguistics. 2016: 225-230.

[7] Lin Y., Shen S., Liu Z., et al. Neural Relation Extraction with Selective Attention over Instances[C]//In Proceedings of the Meeting of the Association for Computational Linguistics. 2016:2124-2133.

[8] Irsoy O, Cardie C. Opinion mining with deep recurrent neural networks[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 720-728.

[9] 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). 2016, 1: 919-929.

[10] Cho K., Merrienboer B., Bahdanau D., et al. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[C]//In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014: 103-111

[11] 向晓雯. 基于条件随机场的中文命名实体识别[D]. 厦门大学, 2006.

[12] 朱臻, 孙媛, ZHUZhen,等. 基于 SVM 和泛化模板协作的藏语人物属性抽取[J]. 中文信息学报, 2015, 29(6):220-227.

[13] 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).

[14] 张丙奇, 姜吉发. 企业相关信息抽取技术研究与系统实现[J]. 微电子学与计算机, 2004, 21(1).

[15] 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). 2014: 1556-1567.

[16] Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]//Proceedings of the 14th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1992: 539-545.

[17] DebashisKushary. Bootstrap Methods and Their Application[J]. Technometrics, 2000, 42(2):216-217.

[18] Brin S. Extracting patterns and relations from the world wide web[C]//International workshop on the world wide web and databases. Springer, Berlin, Heidelberg, 1998: 172-183.

[19] Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004: 22.

[20] GuoDong Z, Jian S, Jie Z, et al. Exploring various knowledge in relation extraction[C]//Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005: 427-434.

[21] Lodhi H, Saunders C, Shawe-Taylor J, et al. Text classification using string kernels[J]. Journal of Machine

Learning Research, 2002, 2(3):419-444.
[22] Hasegawa T, Sekine S, Grishman R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2004: 415.
[23] Chen J, Ji D, Tan C L, et al. Unsupervised feature selection for relation extraction[C]//Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts. 2005.
[24] Huang L, Sil A, Ji H, et al. Improving slot filling performance with attentive neural networks on dependency structures[J]. arXiv preprint arXiv:1707.01075, 2017.
[25] Rajani N F, Mooney R J. Supervised and unsupervised ensembling for knowledge base population[J]. arXiv preprint arXiv:1604.04802, 2016.
[26] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.
[27] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[28] Yang Y, Chen W, Li Z, et al. Distantly supervised ner with partial annotation learning and reinforcement learning[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 2159-2169.

作者联系方式：

姓名：马进

地址：江苏省苏州市沧浪区干将东路 333 号

邮编：215006

电话：18351037857

电子邮箱：806379655@qq.com