中国科学: 信息科学 2020 年 第 50 卷 第 7 期: 1003–1018
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
领域大数据知识图谱专题 · 论文

基于简介文本的中文人物关系图谱属性补全与纠错

杨一帆, 马进, 王海涛, 何正球, 陈文亮*, 张民

苏州大学计算机科学与技术学院, 苏州 215006
* 通信作者. E-mail: wlchen@suda.edu.cn

收稿日期: 2019–11–01; 修回日期: 2020–01–18; 接受日期: 2020–03–23; 网络出版日期: 2020–07–13

国家自然科学基金 (批准号: 61525205, 61876115) 资助项目

摘要 一个准确丰富的人物关系图谱不仅能够为大众提供人物实体的清晰介绍和人物之间的相互关联, 而且能够为智能服务系统提供有效的知识支持. 目前大多知识来源均以百科类表格数据为起点, 在此基础上构建知识图谱. 本文主要描述如何充分利用百科类文本数据构建高质量的人物关系图谱. 为解决表格数据中存在属性缺失和错误的问题, 我们采用模式匹配和深度学习模型相结合的策略从文本数据中自动学习属性值, 进行属性补全和纠错, 有效提高了知识图谱的覆盖率和正确率.

关键词 知识图谱, 人物关系图谱, 属性补全与纠错, 信息抽取

1 引言

知识库 (knowledge base) 是能够系统地将客观事实以及知识整理起来的, 具有一定规则关联的知识片集合 [1]. 而知识图谱 (knowledge graph) [2] 是基于相关知识建立的关系网络, 主要用于表示各个实体的概念属性以及实体之间的关系. 在知识图谱中, 节点 (node) 表示实体或者概念, 边 (edge) 表示节点间的关系或者节点的属性, 两者构成网状知识结构 [3]. 知识图谱既能用于查询和展示, 亦能给智能服务系统如 KBQA 等提供有力的知识支持 [4].

常见的知识图谱包含通用领域的知识, 一般注重其整体全面性, 但较少关注人物之间关系. 人是整个社会的主体, 人们通过各自关系进行关联形成人类社会, 因此人物间关系是非常重要的信息. 基于此, 我们希望构造出专注于描述人物实体及关系的知识图谱.

目前知识图谱的知识大多来源于百科类表格数据等半结构化数据. 这些数据数量庞大, 具有非常丰富的信息. 但由于缺乏审核校验, 表格数据常常存在错误或信息缺失, 如何提高图谱知识的覆盖率和正确率成为我们需要解决的主要问题. 本文以多种百科类表格数据作为基础, 定义人物的属性及人物之间的关系. 我们通过对文本进行自动分析抽取相关知识, 实现对知识图谱的补全和纠错, 最终提高了人物关系图谱的属性覆盖率和正确率.

引用格式: 杨一帆, 马进, 王海涛, 等. 基于简介文本的中文人物关系图谱属性补全与纠错. 中国科学: 信息科学, 2020, 50: 1003–1018, doi: 10.1360/SSI-2019-0243
Yang Y F, Ma J, Wang H T, et al. Attribute identification for Chinese inter-personal relation knowledge graph based on encyclopedic text (in Chinese). Sci Sin Inform, 2020, 50: 1003–1018, doi: 10.1360/SSI-2019-0243

本文主要贡献点在于: (1) 从简介文本中识别人物实体属性, 并利用识别结果对中文人物关系图谱进行补全与纠错, 极大地丰富了人物关系图谱的数据来源; (2) 根据属性特点分为常规属性与别名属性, 将常规属性识别转化为序列标注问题, 采用多种不同序列标注模型进行实验, 而别名属性则采用 Bootstrapping [5] 与分类器相结合, 均取得了不错的效果; (3) 将中文字符信息以五笔编码的形式与序列标注模型结合, 提升实验性能.

2 相关工作

目前领域内存在各种类型的知识图谱, 根据领域和用途的不同可分为语言知识图谱、常识知识图谱、领域知识图谱和百科知识图谱等 [6]. 从最初的人工手动构建到目前自动抽取信息, 知识图谱的覆盖范围和包含的知识规模也在不断扩大.

传统手工构建的知识图谱主要基于专家知识, 例如英文的 WordNet [7] 以及中文的 HowNet [8] 等, 其数据质量有所保障, 但是该类图谱无论是覆盖范围或是知识规模都难以达到实用程度. 目前国外知识图谱例如 DBpedia [9, 10], YAGO [11, 12], 以及 Freebase [13, 14], 均以维基百科 (Wikipedia) 等用户生成内容 (UGC) [15] 作为知识来源, 对数据条目组织整理, 建立具有结构化数据的知识图谱. DBpedia 是一个多语言的知识库, 通过制定一系列的结构化知识获取规则从维基百科中摘取事实信息, 并以机器可读的形式存储知识; YAGO 则汇集了 Wikipedia, WordNet 与 GeoNames 等多种数据源, 同时对 WordNet 的词汇定义与 Wikipedia 的分类体系进行融合, 实现更丰富的实体分类体系; Freebase 由元数据构成, 通过类型定义结构化实例, 将数据作为点存储于图数据库中, 并以由源点、关系、目标点、源点值构成的点边数组对数据进行建模. 而国内比较知名的中文通用知识图谱如上海交通大学的 zhishi.me [16] 以及复旦大学的 CN-DBpedia [17] 等, 它们同样是以中文百科类网站起步. 作为当时第 1 个发布的大规模中文语义数据, zhishi.me 对后续工作具有非常重要的影响; 而 CN-DBpedia 经数据过滤、数据融合、知识推断等操作, 形成高质量结构化数据, 目前应用于问答、医疗等各大场景. 随着机器学习和信息抽取技术的发展, 自动构建知识图谱也成为目前一大趋势, 如 NELL (never-ending language learner) [18], 定义了基本的语义关系和数据类别, 从网页中获取海量自由文本, 实现不断迭代获取新的知识.

现有的中文人物关系图谱包括微软人立方关系搜索以及搜狗人物知识图谱等. 人立方在超过十亿个网页中提取人名、地名、机构名等中文实体, 通过算法自动计算相互连接的可能性, 能够描绘复杂的人际关系结构. 搜狗人物知识图谱侧重于对人物实体的介绍, 以及经过整理归纳后的人物属性及相互间的关系. 该图谱以亲情、友情、爱情作为人物关系的 3 大类, 使用户对查询人物所处的社会关系一目了然. 然而上述人物知识图谱均存在一些不足: 人立方知识获取时并不考虑来源的可靠性, 只要文本中存在能够体现关系的知识, 就会被提取添加至系统. 搜狗人物知识图谱则偏向于采用知名度较高的人物实体进行构造. 例如, 娱乐圈存在多个名为 “李晨” 的男明星, 但在搜索过程中, 始终只显示知名度较高的 “李晨” 相关信息.

知识图谱的知识补全与纠错基本来源于外部知识, 通过从其他知识来源获取新的 RDF [19] 三元组来对当前知识图谱进行知识补充. YAGO 采用 WordNet 中的本体知识以及设计字符规则去 Wikipedia 中挖掘新的知识, 并认为每条知识应存在一个置信度 $c \in [0, 1]$, 当获取到的 RDF 三元组已存在时, 增大该条知识的置信度 c; CN-DBPedia 同样采用远程监督结合实体链接等方式从自由文本中获取新的知识, 并提供界面给用户浏览知识, 汇总用户的错误反馈, 以众包的方式来修正知识, 完善当前知识图谱.

本文的人物关系图谱基于可靠的数据来源, 为每一人物实体分配唯一 ID, 利用人物的属性对同名

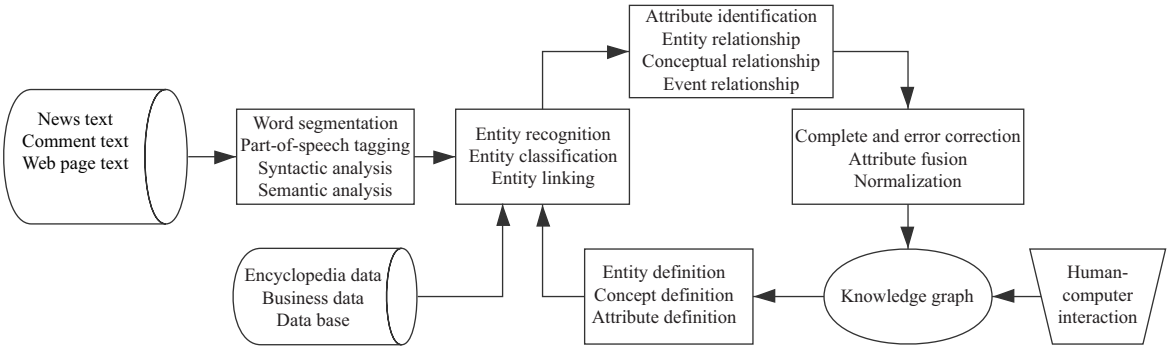

图 1 人物关系图谱构造流程图

Figure 1 Flow chart of inter-personal relation knowledge graph construction

表 1 实体 “姚明” 抽取结果

Table 1 Extraction result of entity “Yao Ming”

Entry ID	Entry name	Entry information (InfoBox)					Entry label
		中文名	外文名	...	生肖	星座	
28	姚明	姚明	Yao Ming	...	猴	处女座	运动员, 篮球, 体育人物

实体进行区分. 同时, 我们利用人物词条简介文本对表格数据进行补全纠错, 取得了较好的效果.

3 人物关系图谱的数据抽取

图 1 描述了整个人物关系图谱的构建流程. 本文侧重于初始数据构建部分, 以百科类数据起步, 经过一系列处理后图谱初具雏形. 后续将通过对文本自动抽取分析等方法, 获取新的事实知识, 不断丰富图谱.

3.1 源网页解析与抽取

我们通过爬虫获取了最新版本的百科词条数据. 常规的词条包含以下几类基本内容: 词条 ID、词条名称、词条简介、词条信息 (InfoBox)、词条标签等. 为构建规范的知识库, 我们着重对基本结构进行抽取, 其他信息选择性抽取. 以运动员 “姚明” 为例, 抽取结果如表 1 所示. 对于每条属性或关系, 彼此存在映射关系 [20], 均能够以形如 ⟨ 姚明, 生肖, 猴 ⟩ 的 RDF 三元组形式表示. 数据经去重和舍弃错误抽取的实体, 最终得到实体约 893 万条, 三元组约 3828 万条.

3.2 人物实体抽取

本文目标旨在构建人物关系图谱, 而从百科网页中获取到的实体类型繁多, 因此需筛选出人物类别的实体. 我们综合词条标签及基本属性, 通过统计所有实体词条信息, 筛选出常见的人物实体属性和标签. 表 2 列举了几种高频人物属性及标签. 统计结果可以看出, 与人物实体关联的属性及标签数量可观. 本文利用两者共同筛选并过滤非人物实体, 最终获得人物实体数量约为 111 万. 我们对筛选结果随机采样 500 条进行评估, 人物词条数量占比为 99.80%.

表 2　人物实体常见属性及标签

Table 2　Common attributes & labels of persons

Attribute	Frequency	Entry label	Frequency
国籍	735080	人物	781199
职业	466398	政治人物	167416
民族	421832	学者	102575
性别	314391	体育人物	97445
毕业院校	228019	娱乐人物	71672

表 3　“出生日期”的多样性

Table 3　Diversity of “date of birth”

Source	Frequency
出生日期	62038
出生时间	21462
出生年月	12227
生日	8552

4　人物属性与关系归一化

Schema 即特定领域下的数据架构, 定义人物数据架构需要规范人物的属性和关系, 有助于知识的标准化以及检索等后续处理. 由于百科数据大多为用户生成内容, 编辑者可能会使用不同的名称格式表达同种属性或关系. 因此为了让图谱拥有统一的属性和关系名称, 进行归一化操作是有必要的. 归一化 (normalization) 是指通过对数据的处理, 将数据限制在自身需要的设定范围以内. 在本文中, 归一化操作包括属性归一化和关系归一化.

4.1　人物属性归一化

人物属性, 是人物实体本身的概念. 如 “张三的性别是男”, “性别” 即 “张三” 人物实体的属性. 对于人物属性, 其归一化包括属性名称归一化和属性值归一化. 属性名称归一化是将表达相同属性意义的多个字串归并为同一字串. 以 “出生日期” 为例, 它是人物实体出现频次较高的属性之一, 且出现形式多样. 表 3 列举了表达 “出生日期” 的几种字串, 我们令这些字串均归并为 “出生日期”. 对于属性值的归一化, 归纳出常见的值类型, 并根据每一类别总结出合理的归一化格式. 部分示例如表 4 所示.

4.2　人物关系归一化

人物实体之间关系同样需要归一化. 我们对人物关系进行归纳, 整理出词条中包含的人物关系列表, 例如 “丈夫”、“女儿” 等, 其对应的值应为另一人物实体. 对于表达同一关系的不同字串, 我们均进行归一处理. 图 2 列举了几种表达 “老师” 的不同字串. 根据当前的人物实体知识库, 能够统计得到 264 种不同人物关系字串. 将这些关系归一整合, 最终我们得到常见人物关系 63 种. 我们把这些人物关系整理成人物关系树, 受篇幅所限, 仅列举出如图 3 所示的部分关系.

在爬虫抓取过程中, 依据词条信息中属性值带有的超链接, 实现网页跳转, 利用归一化后的关系属性将跳转前后的人物实体连接, 构建了初步的人物关系图网络. 接着通过整理后的关系逻辑进行基本的关系补全, 例如实体 A 拥有 “父亲” 实体 B, 我们便可以根据实体 A 的 “性别” 属性来补全实体 B

表 4 属性值归一化示例

Table 4 Examples of attribute value normalization

Type of value	Original format	Normalized format
Time type	1992.9.9/1992-9-9 /1992 年 9 月 9 日	1992-09-09
Number type	182 cm/182 CM/182 公分	182 厘米
	56 kg/56 Kg/56 公斤	56 千克
	1.82 m/1.82 M	1.82 米
String type	1 米 82/一米八二	1.82 米
	「红楼梦」/【红楼梦】	《红楼梦》

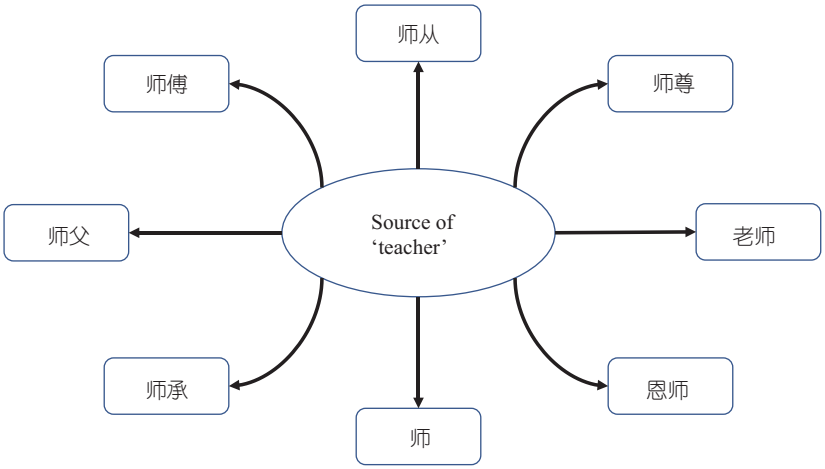

图 2 (网络版彩图) 表达 "老师" 关系的字串

Figure 2 (Color online) Expressions related to "teacher"

中 "儿子" 或 "女儿" 关系的值. 未来工作将专门针对人物实体间的关系进行补全与纠错.

5 属性补全与纠错

经过上述获取步骤, 得到了基本的人物实体数据集合. 但是这些数据面临两大主要问题: (1) 属性值缺失; (2) 属性值存在错误. 针对这两个问题, 我们将人物属性分为常规属性及别名两类, 基于词条简介文本实现人物属性补全和纠错.

5.1 常规属性识别与纠错

本小节主要对词条简介文本进行属性识别, 如表 5 所示, 即实体 "姚明" 的词条简介文本. 该文本包含许多有用知识, 其中一些是表格数据未覆盖到的. 基于此, 我们希望从词条简介中识别出人物实体的属性并回填至人物关系图谱中. 根据属性的不同特点, 我们将常规属性识别分为两类: (1) 基于规则的属性识别; (2) 基于模型的属性识别.

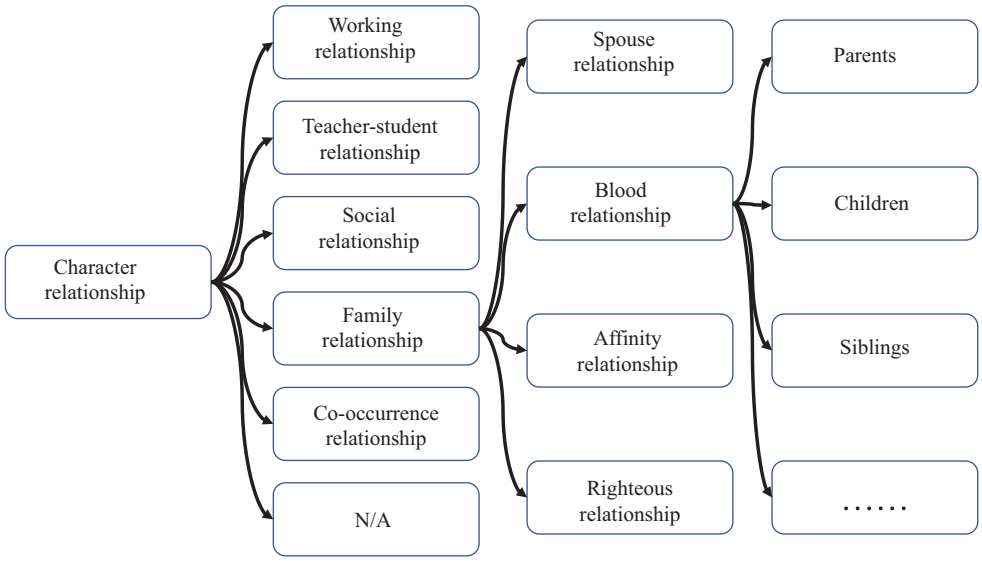

图 3 (网络版彩图) 部分人物关系树

Figure 3 (Color online) Part of the inter-personal relation tree

表 5 词条简介文本示例

Table 5 Example of entry introduction text

Entry ID	Entry name	Entry introduction text
28	姚明	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 ...

表 6 “身高” 属性值识别的部分模式示例

Table 6 Some pattern examples for identifying the value of “height”

Pattern	Example
[1-2]{1}[0-9]{1,2}(cm|CM|Cm|cM|厘米|公分)	183 cm/183 CM/183 公分
[1-2]{1}(\.[0-9]*)?(米|公尺|m)([0-9]*)?	1 米 83/1.83 m/1 米/1 m
(([一二]|[1-2]){1})(米)([一二三四五六七八九]{1,2}|[0-9]*)	一米八/1 米八三/一米 83/1 米 83
([一二三四五六七八九])(尺|英尺)([一二三四五六七八九]{1}(寸|英寸)?)?	一英尺三英寸/三尺八

5.1.1 基于规则的属性识别

若人物实体某属性对应的属性值具有明显规律, 对于此类属性, 我们逐一设计用于识别该属性值的模式, 例如 “身高” 属性值识别的模式如表 6 所示. 针对识别出的属性值, 还需与属性进行关联, 例如加入相关字符限制等关联规则: 若识别结果的上下文出现字符串 “身高”, 则将该结果作为身高的属性值. 该做法可提高正确率, 同样也可能过滤掉一些潜在的正确答案, 因此最终以实验结果来评估关联规则的好坏.

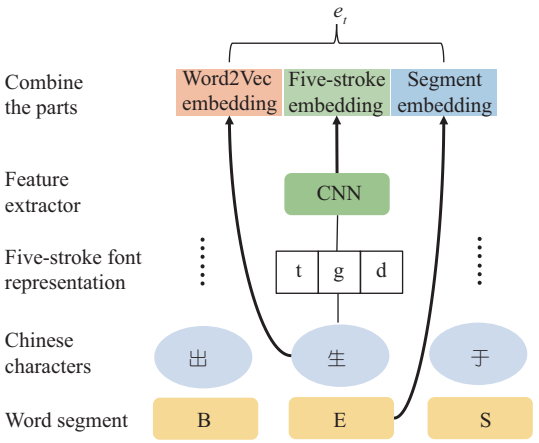

图 4 (网络版彩图) 字向量表示方法

Figure 4 (Color online) Char vector representation

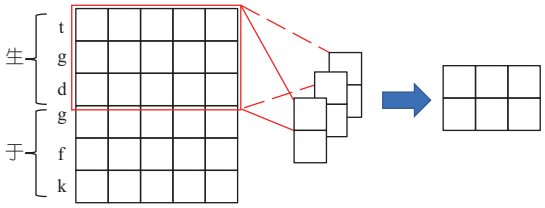

图 5 (网络版彩图) 基于 CNN 的五笔编码特征提取

Figure 5 (Color online) Feature extraction of Five-Stroke coding based on CNN

5.1.2 基于模型的属性识别

基于模型的属性识别, 我们将其转化为序列标注问题. 目前有许多可用于解决序列标注问题的模型 [21~25]. 经初期实验结果比较, 最终采用基于五笔编码的 FS-BiLSTM-CRF 序列标注模型构建属性识别系统.

同种类型的属性值通常有相似的字形笔画, 如 "运动项目" 的属性值集合中, "足球" 的 "足"、"跳高" 的 "跳"、"赛跑" 的 "跑" 等, 这些字符有相似的字形特征. 因此, 本文在 LSTM 基础上, 引入基于五笔编码的字向量表示来加强模型的表达能力. 五笔字型输入法 (five-stroke font input method) 是一种典型的形码输入法. 该输入法依据字形及笔画特征, 对汉字进行编码, 具有重码率低等特点. 通过获取常用汉字对应的五笔编码, 作为词表与模型相结合. 对于能够以五笔编码的汉字, 采用 CNN 模型提取该汉字的五笔编码特征. 对于不在五笔词表内的汉字或非汉字字符, 使用指定字符编码统一表示. 此外, 考虑到分词信息的影响, 我们预先对句子分词, 并将词切分信息编码得到 segment embedding. 最终将它们与 Word2Vec 字向量拼接, 得到新的字向量表示.

$$fs_t = \mathrm{CNN}(\mathrm{FS}(x_t)), \tag{1}$$

$$e_t = w_t \oplus fs_t \oplus s_t, \tag{2}$$

其中 x_t 表示输入序列第 t 个时刻的汉字, w_t 为该汉字的 Word2Vec 字向量, s_t 为词切分信息编码, FS 是将字符转换为五笔编码的函数. 以 "生" 字为例, 最终该字向量表示如图 4 所示, 其中, CNN 模块是表示基于 CNN 的五笔编码特征提取, 如图 5 所示.

LSTM 以 RNN 作为基础, 增加了输入门、输出门、遗忘门, 根据序列信息的重要性进行保留与遗忘, 能够处理序列长依赖问题. LSTM 的主要公式如下所示:

$$i_t = \sigma(e_t \cdot W_{\mathrm{ei}} + h_{t-1} \cdot W_{\mathrm{hi}} + b_i), \tag{3}$$

$$o_t = \sigma(e_t \cdot W_{\mathrm{eo}} + h_{t-1} \cdot W_{\mathrm{ho}} + b_o), \tag{4}$$

$$f_t = \sigma(e_t \cdot W_{\mathrm{ef}} + h_{t-1} \cdot W_{\mathrm{hf}} + b_f), \tag{5}$$

$$\tilde{c}_t = \tanh(e_t \cdot W_{\mathrm{ec}} + h_{t-1} \cdot W_{\mathrm{hc}} + b_c), \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \tag{7}$$

$$h_t = o_t \odot \tanh c_t, \tag{8}$$

其中, e_t 代表 t 时刻的输入, σ 代表 sigmoid 函数, $\odot$ 代表元素点积运算, i_t, o_t, f_t 分别为 t 时刻的输入门、输出门, 及遗忘门的结果, $\tilde{c}_t$, c_t, h_t 分别为 t 时刻的候选细胞状态、细胞状态, 及隐层状态. 对于每一单位输出 e_t, LSTM 层会考虑从 e_1 至 e_t 的上下文信息, 即隐层状态输出 h_t.

CRF [22] 层能够合理化标签序列. 对于一组输入序列 $X = (x_1, x_2, \ldots, x_n)$, 对应预测标签序列 $Y = (y_1, y_2, \ldots, y_n)$, 其得分定义如下:

$$\mathrm{Score}(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}, \tag{9}$$

其中 A 为转移得分矩阵, $A_{y_i, y_{i+1}}$ 表示从标签 y_i 转移至标签 y_{i+1} 的得分, P_{i, y_i} 表示第 i 个输入状态被标注为标签 y_i 的得分, y_0 和 y_{n+1} 分别为标签序列的起始标签和结束标签. 最终经过归一化得到每个标签序列的条件概率:

$$p(y|X) = \frac{e^{\mathrm{Score}(X, y)}}{\sum_{\tilde{y} \in Y_X} e^{\mathrm{Score}(X, \tilde{y})}}, \tag{10}$$

其中 Y_X 表示输入序列 X 所有可能的标签序列集合. 训练时即最大化式 (10) 的真实标签序列的对数似然概率, 最终对自由文本标注时, 取满足下式的结果 y^* 作为最佳预测标签序列:

$$y^* = \arg\max_{\tilde{y} \in Y_X} \mathrm{Score}(X, \tilde{y}). \tag{11}$$

针对属性 p, 首先获取含有该属性的人物实体. 若人物实体包含 p 属性且属性值存在于词条简介中, 标记该词条简介并添加到训练集中, 用于训练针对 p 属性的识别模型. 最后利用训练好的模型, 对所有人物实体识别 p 的属性值.

5.1.3 常规属性纠错

当属性原有值不符合预期时, 我们希望以较小的风险来纠正该类错误, 因此本文属性纠错的主要对象为属性值能够进行归一化的属性集合. 对于属性值能够进行归一化的常规属性 p^*, 找出属性值不满足该归一化格式的人物实体集合 H_{p^*}, 认为该人物集合的属性 p^* 值存在错误. 我们逐一对 H_{p^*} 内人物实体的简介文本进行 p^* 属性值的识别. 若系统识别出结果, 则进行替换, 否则令该实体的 p^* 属性值为空.

我们从网页表格中获取的人物属性存在一些错误, 如存在 “沈勇” 实体, 其 “性别” 属性值为 “汉”, 不满足该属性的归一化格式. 此时认为该实体的 “性别” 属性值存在错误, 若系统能够在简介文本中识别出合适的 “性别” 属性值, 则进行替换, 否则将该人物的 “性别” 属性值置为空.

5.2 别名属性库构建

别名 (alias) 是指官方规范以外的名称, 一个人物实体可以拥有多个别名, 如昵称、爱称等. 例如用户希望能够搜寻 “华仔” 来获取实体 “刘德华” 的信息. 在本文中, 对于任一实体, 我们令其词条名称作为规范名称, 其他能够明确表示该实体且与词条名称不相同的字串, 均为该实体的别名.

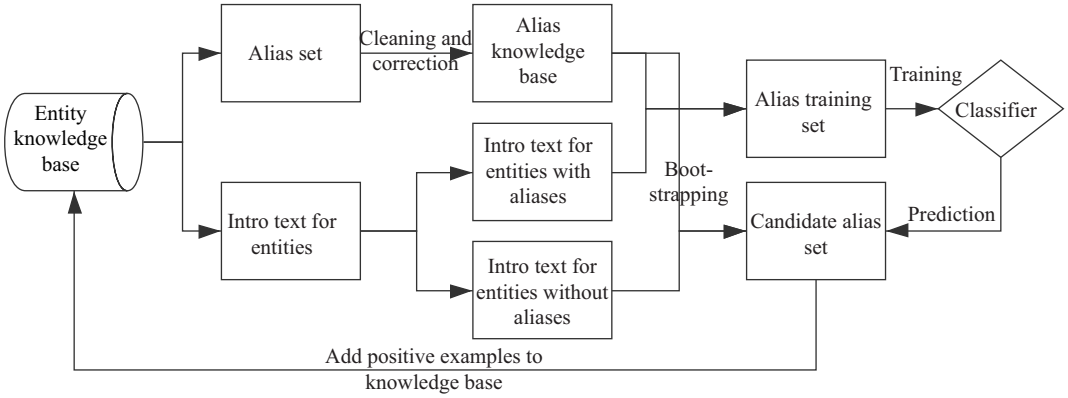

图 6 别名识别模块框架图

Figure 6 Flow chart of alias recognition module

表 7 “别名”常见来源

Table 7 Common sources of “alias”

Source	Frequency	Source	Frequency
中文名	1015465	字号	22960
外文名	163707	其他名称	9451
别名	74988	英文名	9010
本名	43595	别称	7585

别名在属性中具有重要地位, 人物实体别名与当前词条名称两者均能够代表该实体. 此外, 别名能够拥有多个属性值, 其值类型亦多样化, 因此无法采用基于规则的方式进行识别; 而在识别过程中也可能会抽取到非当前实体的别名. 综上考虑, 我们将别名与常规属性进行区分, 对其单独识别处理.

本文采用 Bootstrapping [5] 的方法, 以 “value-pattern-value” 的方式迭代, 获取新的 ⟨ 实体 – 别名 ⟩ 候选对. 对于抽取结果, 我们构建一个分类器来进行自动判别. 最终, 挑选出置信度较高的部分 ⟨ 实体 – 别名 ⟩ 对, 回填至人物关系图谱. 图 6 为别名识别模块的框架图.

5.2.1 种子别名集合构建

利用第 4 节的归一化结果, 我们得到部分能够作为别名来源的属性. 表 7 列举了几种常见的别名表达. 我们令具有别名属性的实体集合为 $\tilde{I}$, 对于人物实体 $s \in \tilde{I}$, 其别名集合为 A_s, 词条简介为 d_s, 则对于任一名称 $a \in A_s$ 且 $a \neq s$, 均为 s 的别名.

由于用户生成内容格式不一, 为方便后续处理, 我们需要对作为种子的别名集合 A_s 进行清洗, 旨在得到每个实体 s 的标准别名集合. 清洗结果示例如表 8 所示. 经清洗, 最终别名集合实体数量为 26 万, 且对任一实体 s 均有 $|A_s| > 0$. 我们对清洗结果人工采样 500 条评估, 正确个数为 496 条, 清洗正确率为 99.20%.

5.2.2 Bootstrapping 抽取

基于高质量的别名集合, 我们寻找满足下列条件的实体及其词条简介: (1) 该实体存在于其词条简介中; (2) 该实体至少有一个别名在其词条简介中.

我们使用正向最大匹配算法 (forward maximum matching method, FMM), 以当前人物关系图谱

表 8　别名清洗结果部分示例

Table 8　Some examples of alias value

Entry name	Alias string	Result after cleaning
亚利克斯 • 维加	阿丽夏 • 维加、Alex 、Lex	阿丽夏 • 维加, Alex, Lex
默森	全名: Paul Mohsen	Paul Mohsen
水莲寺璐珈	水莲寺璐珈 (水莲寺流歌)	水莲寺流歌
观月小鸟	Mizuki Kotori (罗马音) Tori Meadows (英文名)	Mizuki Kotori, Tori Meadows
窦士镛	字晓湘号警凡	晓湘, 警凡

表 9　一般别名的模式抽取

Table 9　Pattern extraction of general aliases

Entry name	Alias	Entry intro
李白	诗仙	S S S 李白 , 唐 代 伟 大 浪 漫 主 义 诗 人 , 人 们 称 之 为 诗仙 . E E E

表 10　附近存在实体或别名的模式抽取

Table 10　Pattern extraction with entities/aliases nearby

Entry name	Alias	Entry intro
泰颉	朱德其, 三友轩主	S S S 泰颉 本 名 朱德其 , 号 三友轩主 , 男 , 1 9 4 0 年 1 月 生 . E E E

表 11　别名抽取模式部分示例

Table 11　Some examples of alias pattern

Pattern	Frequency
, 原 名 ## , 1 9	1504
清] 字 ## , 江 苏	335
\entity , 字 ## , 生 卒	225
名 \alias , ## 等 . E	62

中所有实体和别名作为词表, 对符合上述条件的词条简介进行分词. 对于正确别名, 我们获取其所在位置的上下文, 作为下一轮抽取别名的模式. 为方便模式抽取, 我们在文本首末各添加 3 位特定字符. 模式获取规则如下:

(1) 对于一般的别名, 我们以该别名的前后 3 个字符作为模式. 如表 9 所示: 此时获取到的模式为 “称 之 为 ## . E E”, 其中 “##” 代表需要匹配的字串, “E” 代表句末符.

(2) 当别名周围出现实体或其他别名时, 将其标记并占用一个位置表示它. 如表 10 所示: 当以 “朱德其” 为别名时, 获取到模式 “\entity 本 名 ## , 号 \alias”, 而不是 “颉 本 名 ## , 号 三”. 该过程并不是简单的通过单字特征 (unigram) 获取模式, 而是通过标记 “\entity” 和 “\alias”, 保证该模式在下次迭代时, 相应位置必须对应词条实体和其他别名. 从而在不失准确率的情况下, 提高了别名抽取的召回率.

为获取高质量的模式, 我们对抽取到的模式进行评估. 具体指标包括: (1) 模式出现的频次; (2) 模式在训练集中的正确率. 我们统计每个模式出现的频次以及在训练集中的正确率, 并进行排序, 选取质量较高的模式, 以此进行新一轮别名抽取. 表 11 列举了几种高效准确的模式.

表 12 不同模型在远程监督数据上的实验结果

Table 12 Experimental results of different models on distant supervised data

Model	Precision (%)	Recall (%)	F_1 score (%)
CRF	75.07	80.07	77.49
LSTM-CRF	80.03	80.79	80.41
IDCNN-CRF	80.13	81.40	80.76
FS-BiLSTM-CRF	83.95	81.04	82.47

5.2.3 别名分类器构建

(1) 训练与测试. 对于新获取的别名集合, 我们需要训练一个分类器来自动判别所抽取到的每一〈实体 – 别名〉候选对是否满足别名关系. 我们分别采用最大熵模型 [26], CNN [27], 以及 LSTM [28] 作为分类器模型进行对比实验, 以获取最佳分类效果. 当分类器训练完成后, 我们将分类器认为是正例的结果回填至别名库中, 以此实现从自动抽取、自动判别和自动回填.

(2) 分类特征构造. 别名作为实体的另一昵称, 其指代性较强, 具有明显的特征. 对于大部分别名, 我们能够在文本中找到相关依据及其来源. 本文从别名所在上下文及相对距离构造分类特征.

上下文特征 (contextual feature). 由于别名的指代性较强, 常伴有描述性文本来指定该位置字串为某个实体的别名. 如 “叫作”、“字号”、“名” 等, 较大概率出现在别名周围. 因此列举别名前后多个字符, 构建特征模板.

距离特征 (distance feature). 在没有距离约束的情况下, 可能会抽取到非当前实体的别名, 如 “杜甫, 字子美, 现代主义诗人, 与号称诗仙的李白并称为 ...”. 该词条简介来源于实体 “杜甫”, 但可能会抽取到错误的别名 “诗仙”. 根据观察, 别名与实体距离越近, 该别名是当前实体的别名的概率越大. 因此我们利用两者的距离作为特征.

(3) 训练数据构造. 我们以新获取的别名集合作为测试集, 使用多个模型分别对测试集进行二分类, 依据现有别名集合采用远程监督 [29] 的方法, 并构造正负例. 具体流程为: 遍历模式集合, 对具有别名的人物集合 $\tilde{I}$ 中每一实体 s 的词条简介 d_s 进行模式匹配, 得到 s 的匹配结果集合 $\hat{A}_s$. 已知 s 的别名集合为 A_s, 则 $A_s \in \hat{A}_s$ 且 $A_s = \hat{A}_s - A_s^-$, 其中 A_s^- 是除 A_s 外的匹配结果集合. 对任一别名 $a_s \in \hat{A}_s$, 当 $a_s \in A_s$ 时, 我们将其作为正例; 同样的, 当 $a_s \in A_s^-$ 时将其作为负例.

5.3 实验结果分析

本小节对常规属性识别和别名抽取进行评估, 由于条目数量较大, 且数据集中存在部分正例标签未被标注, 实际模型识别的部分结果被认为是假正例, 导致准确率与实际结果存在偏差. 因此我们采取随机采样的方式进行评估, 对于每条采样记录, 我们经多次鉴定判断正确与否.

为验证本文所采用的序列标注模型的有效性, 我们同时选取了多组不同的序列标注模型进行实验. 我们从基于模型识别的常规属性集合中各抽取 5000 条数据合并作为数据集, 以远程监督的标签结果作为正例集合, 并分别采用了 CRF [22], LSTM-CRF [23], IDCNN-CRF [24], 以及 FS-BiLSTM-CRF 模型进行实验, 最终以 F_1 值衡量各个模型的效果. 不失一般性, 实验采用 5 折交叉验证, 结果如表 12 所示. 从表 12 中可以看出, FS-BiLSTM-CRF 在该数据集上效果最佳, 因此本文最终采用 FS-BiLSTM-CRF 作为属性识别模型.

对于具有明显规律的属性, 采用规则模式对其进行识别. 在表 13 中, 我们列出了部分规则识别结果较好的属性, 其中每个属性各采样 500 条记录进行评估 (共计 3000 条). 而基于模型的属性识别及

表 13　基于规则的属性识别实验结果

Table 13　Experimental results of rule-based attribute recognition

Attribute	Range	Sampling accuracy (500) (%)
身高	Numbers, letters and Chinese	99.20
体重	Numbers, letters and Chinese	96.20
三围	Numbers, letters	95.00
星座	Chinese	96.00
血型	Letters, Chinese	97.40
政治面貌	Chinese	98.60

表 14　基于模型的属性识别实验结果

Table 14　Experimental results of model-based attribute recognition

Attribute	Range	Category	Total number	Sampling accuracy (500) (%)	
出生日期	Numbers, Chinese	A	627454	99.00	
		B	114467	99.00	
		C	72348	100.00	
		D	32856	96.20	93.40
运动项目	Numbers, Chinese	A	64773	96.80	
		B	56743	95.80	
		C	41918	99.80	
		D	716	96.40	95.60
出生地	Chinese	A	62474	97.20	
		B	113837	93.80	
		C	31327	99.00	
		D	20889	97.00	87.20
学位	Chinese	A	51374	93.40	
		B	218215	91.00	
		C	30153	97.40	
		D	21221	87.40	84.60
毕业院校	Chinese, English	A	234951	94.20	
		B	87901	96.80	
		C	10620	99.60	
		D	7475	92.80	89.20
民族	Chinese	A	396757	94.60	
		B	25389	93.00	
		C	14567	99.80	
		D	2362	93.20	88.60

补全纠错, 我们统计如下数据来评估实验效果, 具体如下:

-A: 属性 [**原有值**] 数目及采样正确率;

-B: 属性系统识别 [**识别值**] 数目及采样正确率;

-C: **原有值**和**识别值**都存在且一致条件下, 数目及采样正确率;

-D: **原有值**和**识别值**都存在且不一致条件下, 数目及各自采样正确率.

限于篇幅, 表 14 仅展示部分属性的模型识别效果, 其中我们对每个属性的每一类别 (A, B, C, D)

表 15 属性补全前后覆盖率对比

Table 15 Coverage comparison before and after attribute completion

Attribute	Original coverage (%)	New coverage (%)/△ (%)	Attribute	Original coverage (%)	New coverage (%)/△ (%)
学位	4.61	24.21/+19.60	民族	35.63	37.02/+1.39
出生地	5.61	15.83/+10.22	身高	7.93	8.32/+0.39
出生日期	56.36	66.05/+9.69	体重	5.80	6.06/+0.26
毕业院校	21.10	28.78/+7.68	三围	0.27	0.51/+0.24
运动项目	5.81	10.21/+4.40	星座	2.58	2.65/+0.07
政治面貌	2.80	4.45/+1.65	血型	0.02	0.03/+0.01

表 16 别名抽取的分类实验结果

Table 16 Experimental results of alias extraction

Model	Precision (%)	Recall (%)	F_1 score (%)
MaxEnt	98.07	38.93	55.73
CNN	94.59	44.52	60.55
LSTM	95.85	47.07	63.14

均各采样 500 条进行评估 (共计 12000 条). 从结果可以得出, 识别值与原有值相比整体结果基本一致, 系统可以较好对原有数据进行补全纠错, 但也有部分属性效果需要进一步改进提高. 对于模型识别效果不佳的属性, 我们发现原有值和识别值都存在且一致时 (即类别 C) 具有较高正确率. 因此, 对这些属性只保留原有值和识别值一致的结果. 表 15 为部分属性补全前后的覆盖率对比.

针对人物别名抽取, 本文采用不同模型与人工规则相结合构建分类器进行实验. 我们将 5.2.2 小节中所得模式按照一定条件进行筛选过滤, 保留在训练集中出现频次大于 2 且正确率大于 30% 的模式, 并以此采用多个模型作为分类器模型进行对比实验. 我们对不同模型的分类结果分别采样 500 条, 采用 $P/R/F_1$ 指标人工评估模型分类实验, 实验结果见表 16.

可以看出实验结果的准确率较高, 达到回填至知识库的标准. 其中最大熵模型拥有最高的准确率但召回率不太理想, 综合效果低于深度学习模型; CNN 模型的准确率与召回率均低于 LSTM 模型; 而 LSTM 在拥有较高的准确率同时达到最高的召回率. 我们希望在保证准确率的前提下, 尽可能多地获取别名并回填至知识图谱中, 因此最终我们挑选 F_1 值最高的 LSTM 作为分类器的最终结果, 共抽取到新的 ⟨ 实体 – 别名 ⟩ 对 8908 条, 综合准确率为 95.85%.

6 总结与展望

本文描述了从百科类数据构建人物关系图谱的整个过程. 我们对人物属性和关系进行归一化处理, 将相同意思的不同表达汇总为同一名称. 为提高图谱的覆盖率和正确率, 我们从文本中自动学习知识进行属性补全和纠错. 最终获得包含百万级人物实体和千万级三元组的人物关系图谱.

当前版本的人物关系图谱规模仍然较小, 属性和关系信息缺失仍较严重. 对于不符合归一化格式的属性, 本文对属性纠错任务进行了尝试, 解决了部分问题. 而当属性原有值与识别值难以区分对错时, 需要更加丰富的语料与背景知识支撑, 仅依靠本文的方法难以解决. 我们拟在将来工作中考虑多种不同来源的数据交叉检验, 进行纠错. 更重要的是在人类社会中, 人物关系时时刻刻都在变化. 如何紧跟变化的步伐, 实时更新图谱条目是我们亟待解决的问题. 今后的工作将主要关注知识实时更新问

题, 令系统能够自动阅读新闻, 结合使用命名实体识别系统 [30] 和关系分类系统 [31] 不断获取新的人物实体、属性, 以及人物关系等, 补充至我们的人物关系图谱中.

参考文献

1 Zhong X Q, Liu Z, Ding P P. Construction of knowledge base on hybird reasoning and its application. Chin J Comput, 2012, 35: 761–766 [钟秀琴, 刘忠, 丁盘苹. 基于混合推理的知识库的构建及其应用研究. 计算机学报, 2012, 35: 761–766]

2 Pujara J, Miao H, Getoor L, et al. Knowledge graph identification. In: Proceedings of International Semantic Web Conference, Berlin, 2013. 542–557

3 Liu Q, Li Y, Duan H, et al. Knowledge graph construction techniques. J Comput Res Dev, 2016, 53: 582–600 [刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述. 计算机研究与发展, 2016, 53: 582–600]

4 Cui W Y, Xiao Y H, Wang H X, et al. KBQA: learning question answering over QA corpora and knowledge bases. In: Proceedings of the VLDB Endowment, Munich, 2017. 565–576

5 Abney S. Bootstrapping. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 2002. 360–367

6 Zhao J, Liu K, He S Z, et al. Knowledge Graph. Beijing: Higher Education Press, 2018 [赵军, 刘康, 何世柱, 等. 知识图谱. 北京: 高等教育出版社, 2018]

7 Miller G A. WordNet: a lexical database for English. Commun ACM, 1995, 38: 39–41

8 Dong Z D, Dong Q. HowNet-a hybrid language and knowledge resource. In: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, Toulouse, 2003. 820–824

9 Mendes P, Jakob M, Bizer C. DBpedia: a multilingual cross-domain knowledge base. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, 2012. 1813–1817

10 Auer S, Bizer C, Kobilarov G, et al. Dbpedia: a nucleus for a web of open data. In: Proceedings of International Semantic Web Conference, 2007. 722–735

11 Suchanek F M. YAGO: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web (WWW), New York, 2007. 697–706

12 Rebele T, Suchanek F, Hoffart J, et al. YAGO: a multilingual knowledge base from wikipedia, wordnet, and geonames. In: Proceedings of International Semantic Web Conference (ISWC), Kobe, 2016. 177–185

13 Bollacker K, Cook R, Tufts P. Freebase: a shared database of structured general human knowledge. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence, Vancouver, 2007. 1962–1963

14 Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Vancouver, 2008. 1247–1250

15 Huang R H, Su X B. Development of information organization in social network environment. Libr Tribune, 2011, 31: 190–198 [黄如花, 苏小波. 论社会化网络环境下信息组织的发展. 图书馆论坛, 2011, 31: 190–198]

16 Niu X, Sun X R, Wang H F, et al. Zhishi.me: weaving chinese linking open data. In: Proceedings of International Semantic Web Conference, San Francisco, 2011. 205–220

17 Xu B, Xu Y, Liang J Q, et al. CN-DBpedia: a never-ending chinese knowledge extraction system. In: Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2017. 428–438

18 Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, 2010. 1306–1313

19 Miller E. An introduction to the resource description framework. D-Lib Mag, 1998, 4: 3–11

20 Shen W, Wang J Y, Han J W. Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans Knowl Data Eng, 2015, 27: 443–460

21 Chieu H L, Ng H T. Named entity recognition with a maximum entropy approach. In: Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL, Edmonton, 2003. 160–163

22 Lafferty J, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, Williamstown, 2001.

282–289
23 Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT, San Diego, 2016. 260–270
24 Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, 2017
25 Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. 2015. ArXiv:1508.01991
26 Berger A L, Pietra V J D, Pietra S A D. A maximum entropy approach to natural language processing. Comput Linguist, 1996, 22: 39–71
27 Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, 2014. 1746–1751
28 Wang J H, Liu T W, Luo X, et al. An lstm approach to short text sentiment classification with word embeddings. In: Proceedings of the 30th Conference on Computational Linguistics and Speech Processing, Taiwan, 2018. 214–223
29 Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data. In: Proceedings of Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2009. 1003–1011
30 Yang Y S, Chen W L, Li Z H, et al. Distantly supervised ner with partial annotation learning and reinforcement learning. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, 2018. 2159–2169
31 He Z Q, Chen W L, Li Z H, et al. SEE: syntax-aware entity embedding for neural relation extraction. 2018. ArXiv:1801.03603

Attribute identification for Chinese inter-personal relation knowledge graph based on encyclopedic text

Yifan YANG, Jin MA, Haitao WANG, Zhengqiu HE, Wenliang CHEN* & Min ZHANG

School of Computer Science and Technology, Soochow University, Suzhou 215006*, China*
* Corresponding author. E-mail: wlchen@suda.edu.cn

Abstract An accurate and rich inter-personal relation knowledge graph (KG) not only provides a clear introduction of persons and the interconnections among them but also provide knowledge support for the intelligent service system. At present, most KGs are based on the encyclopedia tabular data. In this article, we mainly describe how to make full use of encyclopedic text to build a high-quality inter-personal relation KG. For solving the problem of missing attributes and errors in tabular data, we propose a method of combining pattern matching and deep learning models to extract attribute information from text data for attribute identification. The experimental results show that our method can effectively improve the coverage and accuracy of KGs.

Keywords knowledge graph, inter-personal relation knowledge graph, attribute identification, information extraction

Yifan YANG was born in 1995. He received his B.S. degree in computer science and technology from Soochow University, in 2017. He is currently a graduate student at Soochow University with major in software engineering. His research interests include natural language processing and knowledge graph.

Haitao WANG was born in 1996. He received his B.S. degree in computer science and technology from Soochow University, in 2018. He is currently a graduate student at Soochow University with major in computer science and technology. His research interests include natural language processing and knowledge graph.

Jin MA was born in 1995. He received his B.S. degree in computer science and technology from Soochow University, in 2018. He is currently a graduate student at Soochow University with major in computer science and technology. His research interests include natural language processing and knowledge graph.

Wenliang CHEN was born in 1977, he received his Ph.D. degree from Northeastern University, in 2005. Currently, he is a professor at Soochow University. His research interests include natural language processing and knowledge graph.