文章编号:1000-5862(2013)03-0279-05

面向非结构化文本的开放式实体属性抽取

曾道建,来斯惟,张元哲,刘　康,赵　军*
(中国科学院自动化所模式识别国家重点实验室,北京　100190)

摘要:从非结构化文本中抽取给定实体的属性及属性值,将属性抽取看作是一个序列标注问题. 为避免人工标注训练语料,充分利用百度百科信息框(Infobox)已有的结构化内容,对非结构化文本回标自动产生训练数据. 在得到训练语料后,结合中文特点,选取多维度特征训练序列标注模型,并利用上下文信息进一步提高系统性能,进而在非结构化文本中抽取出实体的属性及属性值. 实验结果表明:该方法在百度百科多个类别中均有效;同时,该方法可以直接扩展到类似的非结构化文本中抽取属性.

关键词:属性抽取;非结构化;信息框;百度百科
中图分类号:TP 391　　**文献标志码**:A

0　引言

互联网上存在大量的非结构化电子文本,如新闻、博客、电子邮件、政府文件、聊天记录等. 如何帮助人们正确理解这些数据? 普遍的观点是通过注释语义信息,把非结构化文本变成结构化文本. 但是,巨大的数据量以及数据的异质性,使得不可能完全依靠人工来实现这种转换. 迫切需要利用计算机自动地从爆炸式增长的互联网数据中抽取出结构化信息.

目前开放式信息抽取技术多以实体为核心[1]. 实体是指独立存在的事物,不同类型的实体一般具有不同的属性,同一类的实体一般具有大致相同的属性,只是其属性值会有所不同. 例如“人物”类实体一般有国籍、出生日期、职业、毕业院校等属性,“电影”类实体一般有导演、主演、编剧等属性. 非结构化实体属性值抽取是对一个给定的实体,从非结构化文本中抽取出实体的属性及其属性值形成结构化数据. 若给定一个实体 A,将其属性值看作实体 B,属性看作他们间的关系,非结构化属性抽取是实体关系抽取任务,目标是给定实体,从非结构化文本中抽取出(实体,属性,属性值)三元组.

现有的开放式属性抽取工作主要在英文文本上进行,其方法也多为使用人工标注的语料或需要句法分析的结果. 这些方法较难扩展到互联网级别的中文文本实体属性抽取.

本文提出了一种从非结构化文本中抽取给定实体的属性及属性值方法,为避免人工标注训练语料,充分利用网络上已有的结构化内容(百度百科的信息框),对非结构化文本回标自动产生训练语料,同时根据回标的结果自动确定出需要抽取的属性. 本文将属性抽取看作是一个序列标注问题,在得到训练语料后,结合中文特点,选取词形、词性、上下文等多维度特征训练序列标注模型,进而在非结构化文本中抽取出实体的属性及属性值.

1　相关工作

传统的实体关系抽取是给定关系类别,在限定语料中判断 2 个实体是否存在给定的关系,在面对网络级别的信息量时,传统的方法显得无能为力. 随着互联网数据的爆炸式的增长,O. Etzioni 等[2]提出了面向网络信息的抽取系统 KnowItAll. 它通过输入少量实体和关系的实例得到搜索引擎,利用 Bootstrapping 等技术从搜索引擎返回的网页中获得可信的正负样本用来训练分类器,再用该分类器来评估搜索引擎返回的更多网页的抽取结果,获得有限种类的关系及实例. KnowItAll 虽然没有限定领域,但其抽取的关系仍然是受输入关系类别的限制,不能满足从网络文本中获得知识的需求. 在面对海量网络文本资源时,不同的实体类型具有不同关系(或

收稿日期:2012-11-15
基金项目:国家自然科学基金(61070106),国家“973”计划(2012CB316300)和清华信息科学与技术国家实验室(筹)基金资助项目.
通信作者:赵　军(1966-),男,山西晋城人,研究员,博士生导师,主要从事自然语言处理、信息抽取和问答系统的研究.

属性),并且由于海量网络文本的不规范性、开放性,传统限定领域和限定关系类别的信息抽取技术受到人工定义关系类型的限定以及训练语料的限制,很难适应网络文本快速增长、变化的需求. 因此,当前研究热点转向了开放式实体关系抽取.

在开放式实体关系抽取方面,华盛顿大学的人工智能研究组在这方面做了大量代表性的工作,并且开发了一系列原型系统: TextRunner、Kylin、WOE、ReVerb 等. Michele Banko 等[3] 首先提出了开放实体关系抽取,并开发出一个完整的系统 TextRunner,它能够直接从网页纯文本中抽取实体关系. 对于关系名称的抽取,TextRunner 把动词作为关系名称,通过动词链接 2 个论元,从而挖掘论元之间的关系,其抽取过程类似于语义角色标注. TextRunner 首先通过一些简单的启发式规则自动从宾州树库里面获取实体关系三元组的正负样本,根据它们的一些浅层句法特征训练一个分类器来判断 2 个实体间是否存在语义关系; 然后将网络文本作一定的处理后找到候选句子,提取其浅层句法特征,利用分类器判断所抽取的关系三元组是否可信; 最后利用网络数据的冗余信息,对初步认定可信的关系进行评估.

Wu Fei 等还提出了 Kylin 系统,该系统选取包含人工标注 Infobox 的维基百科页面,根据 Infobox 中包含的条目属性及属性值回标产生训练数据,同时根据 Infobox 中的属性名自动确定需要抽取的属性,不同是属性训练不同的 CRF 模型抽取属性值. Wu Fei 等[4] 提出了开放实体关系抽取系统 WOE,该系统首先也是利用维基百科页面 Infobox 回标,通过一些规则挑选含有实体关系的高质量句子,然后使用依存句法分析树的特征以及词性标注浅层特征训练 2 个分类器,作为 2 个实体关系抽取器,以此来获得大量的实体性关系三元组模板. 最后对网络文本的句子做浅层句法处理后,同抽取器获得的模板进行比对,来判断实体关系三元组的可靠性.

Anthony Fader 等[5] 在对 TextRunner 和 WOE 的研究结果进行分析后,根据他们的普遍错误提出了基于句法和词汇约束的实体关系识别器 ReVerb. 它主要解决了以前系统结果中普遍存在实体关系识别错误三元组(Incoherent Extractions) 和无信息量三元组(Un-informative Extractions) 的问题. 实验证明 ReVerb 大幅度提升了关系三元组抽取的准确率和召回率.

由于传统信息抽取方法的局限性,除华盛顿大学外,开放域信息抽取得到许多学者的关注. 文献[6] 研究了怎样在非结构化文本和搜索日志中进行开放域属性抽取,提出了采用弱监督抽取的框架,该方法首先给定初始种子模板,然后通过多次迭代选取置信度高的抽取结果. 在采用 Bootstrapping 进行开放域抽取时,一般初始给定的模板是给定关系找出实例,文献[7-8] 研究了初始时给定一些触发词,通过触发词找出候选的语料,根据候选语料得出元模板,多次迭代后得到最终的抽取模板.

2 非结构化文本属性值抽取

本文以统计模型为核心,对非结构化文本进行实体属性的抽取. 将抽取过程分成 3 部分: 数据预处理、回标产生训练语料、训练统计模型并抽取,算法流程如图 1 所示.

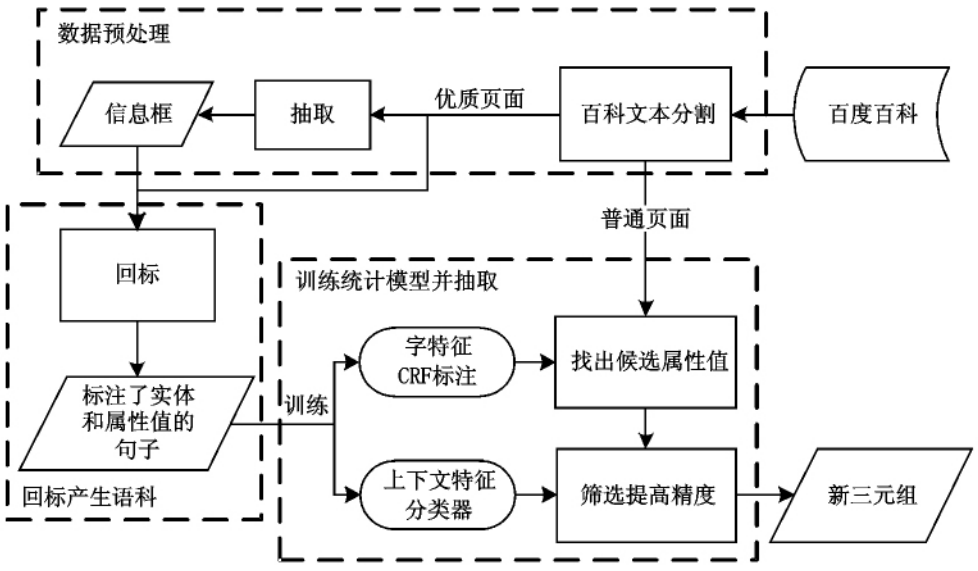

图 1 非结构化属性抽取系统框图

本节以百度百科为例,抽取其非结构化文本信息中的实体属性. 该方法同样可以扩展到其它类似的网页文本中.

2.1　百度百科页面预处理

百度百科词条可以分成词条名、百科名片、信息框(Infobox)、正文文本、开放分类 5 个部分. 从信息组织形式来分,又可将其分为结构化信息(信息框部分)、半结构化信息(部分正文文本)、非结构化信息(百科名片及部分正文文本) 3 种类型.

在百科页面预处理阶段,主要完成以下工作: (Ⅰ) 百科文本分割: 为后续处理方便,将百科页面按期结构分为百科名片等 5 个部分,并分别存储,过滤了页面其它部分(如词条统计、编辑者、广告等信息). (Ⅱ) 去除 HTML 标签、图片等信息,得到纯文本. (Ⅲ) 挑选训练数据: 根据百度百科的"开放分类"将不同的页面分类,然后对各分类训练模型. 如在对"人物"类挑选训练数据时,将百度百科中"开放分类"中包含"人物",并且含有信息框的页面挑选出来. 信息框用于自动标注训练数据,在图 1 中"优质页面"也是指词条中含有信息框. (Ⅳ) 去除"正文文本"的半结构化部分,并进行句子分割. 由于在做非结构化属性抽取时关心的是非结构化部分,因此需要过滤掉百科文本中的半结构化部分. 这部分采用简单的规则实现: 若读入一行不以标点符号结束,就认为这是一个半结构化块. 在挑选出非结构化部分后,将百科名片与正文非结构化部分放在一起,并根据标点符号将文本拆分成句子.

在预处理步骤之后,将原始百科文档转换为词条名、信息框、分句后的词条描述文本 3 部分,并按"开放分类"分类存放.

2.2　回标产生训练语料

统计模型需要大量优质的训练语料作为支撑才能发挥其优势. 面对海量的互联网数据,如果采用人工选取、标注训练语料,将会面临工作量大、时效性差等问题,因此需要设计一种自动方案来完成这一任务.

利用百度百科文档中结构化信息框的内容,对非结构化文本采用自监督学习的方式进行回标,自动产生训练语料,具体步骤如下:

(Ⅰ) 挑选候选句子: 候选的句子必须包括百科词条名或词条名的简称,如"里奥·梅西"词条,候选的句子因包含"里奥·梅西"或者"梅西". 在现阶段工作中,并没用采用指代消解等方式扩充主语是"他"及类似句子,这主要出于 2 个方面考虑: (ⅰ) 指代消解的正确率并不高,使用指代消解会降低候选句子的质量; (ⅱ) 由于网络信息冗余的特点,直接使用词条名及其缩写可以获取足量的候选句子.

(Ⅱ) 遍历信息框内容并回标候选句子: 选取信息框中的各属性值,到(Ⅰ) 中得到的候选句子中逐一匹配,一旦在候选句子中匹配到某个属性值之后,就得到了一句带标注的训练语料.

如果信息框中某条属性的属性值较长并且含有标点符号,首先需要对属性值根据标点符号进行分割,这样一个属性名就对应多个属性值,若候选的句子包含其中任意一个属性,则此句子为这个属性回标的结果. 如在篮球运动员科比词条的信息框中,属性"别名"的属性值为"小飞侠,黑曼巴",则按逗号分成"小飞侠"和"黑曼巴"2 个属性值分别回标.

(Ⅲ) 回标结果过滤: 在(Ⅱ) 中某些文档一个属性回标的结果可能对应多个句子. 选取其中的哪些或哪个句子作为最终回标结果涉及到语义理解的问题. 现阶段根据百度百科的特点,描述实体的文字一般写在开头(如百科名片),故选取第一个回标出来的句子作为最终回标的结果.

在回标阶段,将上一阶段的结果转化为统计模型的训练语料. 训练语料的形式为自然语言文本,并且已经标注了其中的某几个字分别为实体及其属性值,同时也知道这句话是描述哪个属性的. 需要注意的是,属性名不一定出现在句子中.

2.3　训练统计模型并抽取

2.3.1　字特征 CRF 标注　用统计模型抽取实体属性,是利用统计模型自动学习出自然语言中对于某个实体属性的触发词、属性边界等特征.

将从非结构化文本中抽取属性值看着是序列标注问题,而对于序列标注问题,通常采用条件随机场(CRF) 来解决. 将 2.2 节中回标产生的训练语料选取 90% 作为训练数据,剩下的作为测试数据. 对于每一句语料,先对其进行分词及词性标注,然后将词拆分回字,把每个词的词性则对应到该词每个字上,这样就得到了字、所在词的词性这 2 个重要信息. 将前一个字、当前字、后一个字的字及词性的一元、二元组合作为特征训练 CRF 序列标注模型. 对于百科信息框中的每个属性,都对其分配 S、B、M、E 共 4 个标签,其中 S 表示单字属性值; B、M、E 分别表示多字属性值的第 1 个字、中间的若干字及最后 1 个字. 除了属性值对应的标签以外,还有一个 N 标签,用于标示不对应任何属性值的字.

选取字作为特征可以消除对分词性能的依赖,这一策略参照了命名实体识别的相关工作. 由于实

体的属性值经常也是命名实体,选用字作为特征的效果在属性值边界判断上会由于直接用词作为特征的效果.

在具体训练 CRF 模型时,采取了以下 2 种方案.

方案 1:将所有的属性统一训练成为一个模型.在这种情形下,如果打算抽取 n 个属性,则每个属性会对应有 4 个标签,加上 N 标签,一共会有 $4n+1$ 个不同的标签.

方案 2:每个属性训练一个单独的模型.在这种情形下,每个模型有 5 个标签,模型一共有 n 个.

在测试阶段,有可能在一篇文章会抽取出同一属性的多个属性值,考虑到一般百科文本会将概况写在较为前面的地方,规定每篇文章每个属性抽取得到的第 1 个结果作为该实体该属性对应的属性值.

2.3.2 上下文特征分类器 虽然在 CRF 序列标注里已经使用了上下文特征,但是实际上属性值的内部特征占了主导优势.这直接导致了训练得到的模型更多考虑的是某个"属性值"本身是否是一个合理的属性值,而忽略了其是否在上下文语境中讲这个属性.因此,需要在上一步的结果之后,加一个上下文特征的分类器,以提高抽取所得属性值的准确率.

抽出其中标注的属性值,将其替换为一个通配符,同时也将百科的词条名替换成另一个通配符.例如,原文"姚明 出生 于 上海."在字特征 CRF 标注中发现"上海"为属性值,在这里将属性值和词条名替换成通配符后变成"$实体$ 出生 于 $属性$".对于每个属性值训练分类器,取前后 2 个词作为特征.这样就达到了利用上下文语境筛选抽取结果的效果.

3 实验与结果

本文研究对象主要面向中文网络百科,如百度百科、互动百科、中文维基等,主要使用有监督机器学习解决非结构化属性抽取问题,实验验证第 2 节中给出的回标规则的有效性,以及提出的非结构化属性抽取框架的可行性.

在本次实验中,以百度百科为例,选取其中人物和电影 2 个开放分类进行训练和测试.其中对这 2 类各选取了约 1 万个含有信息框的页面,使用 90% 的页面用来训练序列标注模型及上下文分类器,另外的 10% 作为测试集,用于评价方案的效果.在评价中,主要使用准确率和召回率 2 个指标.

3.1 百度百科信息框回标实验

根据 2.2 节的规则得到回标数据后,在"人物"类上随机抽取 20 篇文章用于评价.这 20 篇文章共有信息框属性 165 个,其中有 79 个可以回标出属性值.在这 79 个回标结果中,有 68 个是符合语义的正确回标,另外 11 个虽然属性值的文字匹配正确了,但是在语义上并不对应该属性.在没有回标结果的 86 个属性中,有 53 个在文字本身就没有描述,另外 33 个由于信息框中的属性值与非结构化文本中描述的内容并非逐字匹配而导致没有回标成功.

整个回标结果的准确率为 86.1%,召回率为 60.7%.从实验结果来看,实验的回标方法基本可行,为下一步统计学习提供较高质量的训练样本.

3.2 属性抽取实验

在进行属性抽取实验时,使用回标的结果作为训练语料,并使用 CRF++ 工具包训练 CRF 模型.为了方便评价结果,对"人物"、"电影"类各选取其中回标结果数量较多的若干个属性,训练序列标注模型,并进行人工评价.在实际算法中方案 1 和方案 2 抽取得到的初步结果统计分别见表 1 和表 2.由这 2 个表及表 3 的汇总数据可以看出,方案 1 的准确率高于方案 2,但召回率相比方案 2 低,综合性能略低于方案 2:在上下文特征分类器中会进一步筛选结果,因此在召回率相对较高的方案 2 的基础上,做分类器的实验,实验结果如表 4 所示.实验结果表明,上下文特征分类器在略微降低系统召回率的同时,可以有效地提升系统的准确率.对于电影类,共测试标注了类型、导演、制片地区、主演、编剧、出品时间、上映时间、制片人、出品公司这九类,得到了类似的结果.

表 1 方案 1 抽取人物类属性效果评价

	国籍	出生日期	出生地	职业	运动项目	逝世日期	毕业院校	代表作品	所属运动队	民族
准确率	0.83	0.91	1.00	1.00	1.00	0.50	0.80	0.88	1.00	1.00
召回率	0.65	0.65	0.40	0.37	0.60	0.38	0.38	0.63	0.80	0.33
召回数	17	20	12	14	3	5	8	15	4	2

表 2　方案 2 抽取人物类属性效果评价

	国籍	出生日期	出生地	职业	运动项目	逝世日期	毕业院校	代表作品	所属运动队	民族
准确率	0.77	0.77	0.62	0.76	0.67	0.29	0.71	0.62	0.27	0.50
召回率	0.81	0.77	0.70	0.68	0.80	0.69	0.57	0.67	0.80	1.00
召回数	21	24	21	26	4	9	12	16	4	6

表 3　各方案对人物、电影类抽取时的综合效果评价

抽取方案	类别	准确率	召回率	F 值
方案 1	人物	0.866	0.503	0.636
方案 2	人物	0.608	0.719	0.659
方案 2	人物	0.700	0.683	0.692
方案 1	电影	0.932	0.338	0.496
方案 2	电影	0.608	0.624	0.616
方案 2	电影	0.735	0.484	0.584

表 4　方案 2 + 上下文分类器抽取人物类属性效果评价

	国籍	出生日期	出生地	职业	运动项目	逝世日期	毕业院校	代表作品	所属运动队	民族
准确率	0.85	0.68	0.77	0.93	1.00	0.33	0.73	0.76	0.44	0.56
召回率	0.81	0.68	0.67	0.66	1.00	0.69	0.52	0.67	0.80	0.67
召回数	21	21	20	25	5	9	11	16	4	4

在实验中也发现一些问题,如利用上下文分类器,对于人物类的“逝世日期”并无明显的提升作用. 经分析发现,在人物类的百科词条中,有一种通用的编年体格式“某年某月,某人做了某事”. 逝世日期的回标结果中大量使用了这种格式,但是同时用这种格式描述的事件大多不是逝世,这造成了只用上下文若干词的分类器效果不够明显. 要想改进类似的问题,可以在选取分类器特征时,选取句子中的关键词,而非只找候选属性值附近的词. 在目前的实验中,分类训练时的负样本是随机选取的,如果采用一些规则选取合适的负样本,也必定会有效提升分类器的性能.

4　结论与展望

本文研究了从非结构化文本中自动抽取属性及属性值的方法,利用现有结构化信息自动生成训练语料,通过训练序列标注模型及上下文分类器,能有效地抽取出大量实体属性. 本实验在利用上下文特征时,只使用了最浅层的前后词,并未使用更深层次的分析. 同时对于出现次数较少的属性,统计模型并不能发挥很好的作用. 下一步的工作可以在本文的基础上进行更深层次的上下文特征发现,以进一步提高系统性能; 也可以利用模板等方法抽取出现次数较少的属性.

5　参考文献

[1] 赵军,刘康,周光有,等. 开放式文本信息抽取 [J]. 中文信息学报,2011,25(6):98-110.

[2] Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the web: an experimental study [J]. Artificial Intelligence, 2005, 165(1): 91-134.

[3] Banko M, Cafarella M, Soderland S, et al. Open information extraction from the Web [EB/OL]. [2012-11-12]. http://turing.cs.washington.edu/papers/rjcai07.pdf.

[4] Wu Fei, Daniel S Weld. Open information extraction using Wikipedia [EB/OL]. [2012-11-12]. http://homes.cs.washington.edu/~weld/papers/wu-acl10.pdf.

[5] Oren Etzioni, Anthony Fader, Janara Christensen, et al. Open information extraction: the second generation [EB/OL]. [2012-11-12]. http://turing.cs.washington.edu/papers/etzioni-ijcai2011.pdf..

[6] Marius Pasca, Benjamin Van Durme. Weakly-supervised acquisition of labeled class instances using graph random walks [EB/OL]. [2012-11-16]. http://www.cs.utexas.edu/~joeraii/papers/adsorption_emnlp08.pdf.

[7] Dmitry Davidov, Ari Rappoport, Moshe Koppel. Fully unsupervised discovery of concept-specific relationships by Web mining [EB/OL]. http://www.citeulike.org/user/student_t/article/3270320.

[8] Dmitry Davidov, Ari Rappoport. Unsupervised discovery of generic relationships using pattern clusters and its evaluation by automatically generated SAT analogy questions [EB/OL]. [2012-12-17]. http://www.cse.huj.ac.il/~arir/sat.pdf.

(下转第 305 页)

Journal of Physical Chemistry C,2009,113(32): 14071-14075.

[20] Dale A C Brownson,Craig E Banks. Graphene electrochemistry: an overiew of potential applications [J]. Analyst,2010,135(11): 2768-2778.

[21] Goyal R N,Alok Mittal,Sonam Sbarma. Simultaneous voltammetric determination of hypoxanthine,xanthine,and uric acid [J]. Electroanalysis,1994,6(7): 609-611.

[22] Palraj Kalimuthu,Abraham S John. Simultaneous determination of ascorbic acid,dopamine,uric acid and xanthine usinga nanostructured polymer film modified electrode [J]. Talanta,2010,80(5): 1686-1691.

Simultaneous Determination of Dopamine and Uric Acid on Poly-(*L*-His)/ERGO Modified Glassy Carbon Electrode

JIAN Xuan,YU Hao*,JIN Jun,WANG Yi,LIU Zhen-ye,QI Guang-cai

(College of Chemistry and Chemical Engineering,Yan'an University,Yan'an Shangxi 716000,China)

Abstract: A poly-(*L*-His)/ERGO hybrid film modifying electrode had been fabricated by using cyclic voltammommetry. The electrochemical behaviors of dopamine(DA),uric acid(UA) and ascorbic acid(AA) were investigated. The results showed that the resulting modified electrode exhibited excellent electrocatalytic activity toward the electrooxidation of DA and UA and had superior selectivity for the determination of DA,UA in the presence of mass AA. Under the optimum conditions,the liner calibration curves of $3.0\times10^{-7}\sim3.0\times10^{-5}$ mol·L^{-1} and $5.0\times10^{-7}\sim3.0\times10^{-5}$mol·$L^{-1}$ with the detection is 3.0×10^{-7} mol·L^{-1} and 3.0×10^{-7}mol·L^{-1} were obtained for DA and UA,respectively.

Key words: poly-(*L*-Histidine); graphene; ascorbic acid; dopamine; uric acid

(责任编辑: 刘显亮)

(上接第 283 页)

Open Entity Attribute-Value Extraction from Unstructured Text

ZENG Dao-jian,LAI Si-wei,ZHANG Yuan-zhe,LIU Kang,ZHAO Jun*

(National Laboratory of Pattern Recognition,Institute of Automation Chinese Academy of Sciences,Beijing 100190,China)

Abstract: An approach for extracting attribute-value pairs of a given entity has been proposed,regarding attribute-value extraction as a sequential data-labeling problem. In order to avoid label the corpus manually,the information in the Infoboxes of Baidu encyclopedia is used to label the unstructured text as the training data. After the training data was generated,multidimensional features are used to train the sequential data-labeling model,and then the performance is improved by using the context. Experiments shows that this method can be used in many classes of the Baidu encyclopedia,and this method can be also used in other websites.

Key words: attribute-value extraction; unstructured text; Infobox; Baidu encyclopedia

(责任编辑: 冉小晓)