高技术通讯 2014 年 第24 卷 第6 期:602 ~608

doi:10.3772/j.issn.1002-0470.2014.06.008

基于维基百科和条件随机场的领域主题词抽取方法①

齐保元②[*,**] 史忠植[*]

([*]中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

([**]中国科学院大学 北京 100049)

摘 要 针对传统的手工整理主题词表的方法因耗时、更新速度慢而无法满足应用要求的问题,提出了一种基于维基百科(Wikipedia)和条件随机场(CRF)的领域主题词抽取方法。该方法根据特定领域现有主题词的构词特点、统计分布特点,充分利用维基百科独特的结构优势,自动地从维基百科中获取新的领域主题词,补充现有主题词表。该方法采用条件随机场作为训练、测试的模型,将多方面的特征进行综合建模,取得了较好的实验效果。实验结果表明,使用该方法可以将主题词识别的 F 值提高到83%左右。

关键词 主题词表构建,主题词抽取,维基百科(Wikipedia),条件随机场(CRF)

0 引 言

主题词表也称叙词表(thesaurus),是一种控制词汇的方式,它通过收集特定学科领域的词汇,并以特定的结构排列,以显示词汇之间的关系。主题词表的编制主要包括准备工作、词汇选择、词汇整理、词汇分类、词间关系建立等步骤[1]。传统的主题词表的构建主要采用领域专家手工完成这一方式,耗时较长,无法保证完全覆盖,而且无法进行有效的自动更新,对于新词、组合词、外来词等的接收较慢,因而无法满足实际应用的需求。但我们可以利用维基百科(wikipedia)[2]提供的词条来丰富原有词典中的词条。维基百科是一个由全社会参与的具有多种语言的百科全书协助计划,其目的是建立一个完整正确的百科全书。截至到2012 年2 月底,其收录的英文文章数超过386 万,收录的中文词条接近40 万。维基百科包含社会、经济、文化、教育、科技诸多领域的知识,由具有一定相关领域知识的社区积极分子进行维护与更新,因此对于 Web2.0 带动下的全民织网和进行特定领域下的主题词抽取具有潜在的意义。在维基百科中,各个条目会包含很多的链接,引导用户进入相关的页面,查看更多的词条信息。一个词条包含该词条属于的主条目,而类别条目还会包含该分类的子分类、该分类对应的页面以及参考资料等。维基百科词条的编辑遵循一定的原则,要求词条内容具有正确性、客观性,词条形式具有一定逻辑性,符合其规定的发布规范,扩展内容要能辅助用户了解主题的内涵与外延,从而丰富读者的知识。反过来,作为编辑人员,可以利用原有词典中的词条,在维基百科中建立新的条目,在使用维基百科知识的同时,还可以通过维基百科来分享自己的知识。

本文的工作主要针对前者而展开,从海量 Web 中获取专业领域的词表。我们需要面对信息量的急剧暴涨带来的信息湮没,以及无结构化、半结构化的文本给计算机处理带来的巨大的挑战,因此,如何采用自动、半自动的方法和通过机器学习的技术来解决自然语言理解与自然语言处理中的难题的任务已经迫在眉睫。本文的特色主要是综合了构词特点、统计分布以及维基百科的语义表示符号,将这些特征统一到一个整体模型中。构词特点主要是通过对原有词表中的词汇进行分析,获取常用于组成主题词的前后缀、词性等特征;统计分布是为了计算两个词元在大规模语料中的互信息,这对于是否成词具有很大的统计学意义;利用维基百科特殊的标记符号,有助于我们在进行词汇识别时,增大成词的概率。

本文重点介绍了主题词抽取的相关工作以及我

① 973 计划(2013CB329502),国家自然科学基金(61035003, 60933004, 61202212, 61072085),863 计划(2012AA011003),国家科技支撑计划(2012BA107B02)和中国信息安全测评中心(CNITSEC - KY - 2012 - 006/1)资助项目。

② 男,1985 年生,博士生;研究方向:人工智能,数据流挖掘;联系人,E-mail: qiby@ics.ict.ac.cn

(收稿日期:2013-10-29)

们进行主题词抽取的模型——条件随机场(conditional random field, CRF),描述了实验的流程图和实验中使用的特征及构造方法,给出了实验数据集和实验结果。

1 相关工作

1.1 主题词抽取的相关工作

在电子工程[3]、电子政务[4]、军事训练[5]、电子邮件过滤[6]等领域,都针对特定领域的词表自动构建进行了广泛的研究。词表的自动构建是一种无监督或者半监督的机器学习方法,利用统计特性来发现成词规律,将获得的主题词填入原有的词表中,从而丰富原有的词表,有利于保持特定领域的词典与时代同步、加快信息的聚集速度。

Hearst[7]提出了在自由文本中自动获取下位词的方法,通过一系列诸如 NP0 such as {NP1, NP2, …, (and|or)} NPn 的词汇-文法模式,进行词表的扩充,并发现新的模式。文献[8]研究了基于文法关系的自动词表构建方法,提出了一套客观评价标准,该研究运用文法之间的组合关系,取得了较好的实验结果。然而,这种方法依赖于文法结构,需要人为地维护这些规则,而对于某些特殊的规则需要进行添加或者剔除。

文献[9]将主题词表的构建分为共现分析、概念空间以及贝叶斯网络等三类,并对每类方法进行了分析。Park 等人[10]提出了采用基于 sigmoid 贝叶斯模型的主题词自动构建方法,通过 collocation map 来计算词汇-文档相似度,克服了数据稀疏的问题。然而,该方法要求数据值必须遵循 sigmoid 分布才能在贝叶斯结构中表现出来,具有一定的局限性。Wang 等人[11]提出了在英文/中文平行语料中使用基于 Hopfield 网络的方法,自动构建主题词,具有较高的正确率和召回率。该方法需要有双语平行语料库,在某些条件下不是很现实。

Hagiwara[12]提出了基于 PLSI 的自动主题词构建方法,该方法不仅可以在浅层,而且还可以在深层充分地挖掘隐藏的同类别词汇。Tseng[13]提出了从文档中提取关键词,然后进一步的过滤来进行词汇关联分析,该方法可以根据获得的权重进行关联度排序,并可以反映词汇的时间维的信息,对于研究词汇的演化有较大的辅助作用。王石等[14]提出了一种自动从英文 WordNet 翻译成中文概念的方法,采用基于语义项的对比翻译方式,然后采用分类的方法,获取正确的中文概念词。

曾建勋等[15]提出在网络环境下利用"基础词库-范畴体系-概念关系网络"三级联动机制的主题词表的构建机制;曾文[16]采用基于语言学规则和统计计算的方法建立主题词集合,并通过优化算法进行主题词的选定,提出了建立主题词自动构建的平台与系统;叶春蕾等[17]提出利用文本挖掘技术,并结合科学计量,自然语言处理等方法,提出一种基于三重共现算法的技术路线图中未来技术词表构建方法。

1.2 条件随机场

条件随机场(CRF)是 John Lafferty 于 1993 年提出的判别式模型,在文本处理[18-24]、生物信息学[25-27]和计算机视觉方面[28-30]都取得了较好的实验效果。

令 X 表示观察值集合,Y 表示标注集合,则有

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = (1/Z(\boldsymbol{x}))\exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\right) \tag{1}$$

Z 是归一化因子:

$$Z(\boldsymbol{x}) = \sum_{\mathbf{y}} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \boldsymbol{x}_t)\right\} \tag{2}$$

其中,$f_k(y_t, y_{t-1}, \boldsymbol{x}_t)$ 为用户自定义的特征函数,表示观察序列 x 中位置为 t 和 $t-1$ 的输出节点的特征,λ_k 为每个特征函数的权值。通过 L-BFGS 等算法,可以对模型参数进行训练。

1.3 系统框架

本文提出的主题词抽取方法是建立在一系列的数据转化之上的,整个流程尽量减少人工的参与,特征抽取过程中,完全不需要人工参与;在人工标注过程中,为了减少标记可能带来的错误,采用了可视化标注的方法。

主题词抽取的整体流程如图 1 所示。

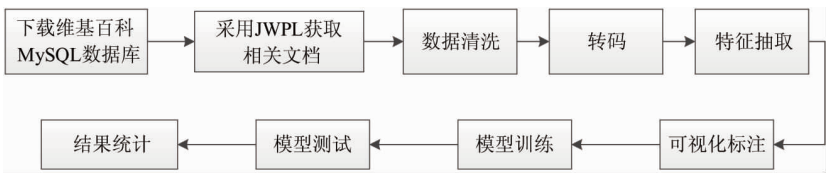

图 1 主题词抽取的流程图

（1）维基百科提供了中文数据的 MySQL 数据库文件，有全备份以及在此之上的增量备份，这样我们操作数据的自由更大，不必受限于其官方网站提供的 API 的调用频度限制。

（2）JWPL（Java WikiPedia Library）[31]是一个基于 Java 的类库，提供了对维基百科 MySQL 数据库的各种操作，通过提供的 API，我们可以对所有的数据进行透明的访问，包含了查询所属分类、子分类、页面的数据、页面的标题等功能。

（3）数据清洗是为了去掉文档集中与所选文档集合无关的文档，提高数据集的质量。

（4）特征抽取采用了 ICTCLAS[32] 与 WikipediaTokenizer[33] 相结合的方法。首先使用 WikipediaTokenizer 对文档集进行语义标记的分割形成单独的块，然后对这些块采用 ICTCLAS 进行分词与词性标记，在此基础上抽取长度、是否是分割标记等相关的特征。互信息的抽取则需要在对全部文档进行扫描的基础上单独进行。

（5）可视化标注主要是为了解决人工标注的不方便。我们通过提供标注的页面，用户只需要在标注时进行划词操作，选中可以构成主题词的连续字符，然后，我们对用户的标注结果进行处理，生成用于条件随机场的序列标签。

（6）模型的训练与测试采用CRF++[34]工具包进行。模板设置的窗口长度为 3。

本文主要的工作在于 4 和 6 两大模块，在下一节将对特征的构造进行详细的阐述，而训练和测试在第 4 部分的实验中。

2 条件随机场特征的构造

特征的选取对于条件随机场具有决定性作用，我们选取特征主要考虑了如下几个方面的因素：

（1）词本身的特征，比如是否具有常用的前后缀、词性、长度，是否是数字，是否是分割标记等，从词语构成的角度来分析，具备常用前后缀、名词性、长度适中、非数字、非字符分割标记等特征的词语更具有组成主题词的可能性，因此，这些特性对于我们模型的训练有较大的提升作用。

（2）词在文档中的上下文信息，主题词的形成需要在一定的语境环境中出现，比如，形容词或者动词后容易跟着主题词，多个名词容易形成组合的主题词，前后两个词的互信息满足一定的阈值条件就可以合并等，这些上下文的环境对于主题词的识别会有较大的帮助作用。

（3）词在维基百科中的标记信息，人们在维基百科中编辑词条时会采用约定的标记方法，例如加粗显示、是否是一个分类，是否是一个内部链接、是否是一个主要的段落标记头信息等，这些是描述一个主题词的重要的信息，我们在对语料加工时可以将这些信息作为特征提取出来，以提高主题词识别的效果。

下面详细介绍我们在条件随机场中使用的特征。

2.1 原始词表的组词特点

对原有词表中的词条进行分析，可以发现很多词具有较为明显的组词规律。比如“军用飞机”、“水上飞机”、“垂直起落飞机”等条目，具有相同的后缀词“飞机”，而“航空电子学”、“航空地图”、“航空材料”等条目，具有相同的前缀词“航空”。我们把这些可以当做常用后缀或者常用前缀的词语称为“常用词元”，并把常用词元区分为前缀词元和后缀词元两种。有些词元可能既是前缀词元也是后缀词元，比如“地图”、“发射”等。

我们进行的实验主要考察词元的分布规律，因此首先要进行词元的提取。本实验定义的词元是具有一定的成词能力并具有一定统计优势的词语构成部分。本实验采用的方法是结合最长前（后）缀提取的。

实验中选取实验阈值为 4，即当公共的前缀或者后缀的出现个数大于等于 4 时，则作为一个常用的前后缀词元。这需要借助于前缀树来实现。

对于单词集合 S，构建前缀树的算法如算法 1 所描述。

算法 1 构建前缀树的过程

```
输入：单词集合 S
初始化：建立空树 T = Ø
while 单词集合 S 不空
    从 S 中任取一个元素 I
    if  T = Ø
        则将元素 I 作为 T 的根节点；
    else
        在 T 中查找与之有公共最长前缀的节点
        if 无法找到最长前缀节点，
          直接插入到根节点；
          else 将原来的节点分裂为公共最长前缀与剩
               余部分组成的节点，并将 I 的剩余部分作
               为孩子节点插入到该公共最长前缀上。
```

```
        end if
    end if
    从 S 中删除元素 I
end while
输出:前缀树 T
```

对应的后缀树的构造过程与此相似,只需要将原有的词语进行倒置,然后使用该算法即可构建一颗后缀树。

构建好前(后)缀树以后,在对树的挖掘中,将孩子总数超过阈值的节点提取出来,即可作为常用的词缀。

最终提取出的常用词元共有 264 个,表 1 是抽取出的最常见词元的举例。

表 1 "太空"领域主题词的常用词元(部分)

回收	倾斜	撞击	测量
外挂	混合	发射	监视
助推	扭转	屏蔽	起飞
轰炸	爬升	设计	跟踪

表 2 是部分常用词元中在原有词表中的词首、词中、词尾的分布情况。

表 2 常用词元在词表中的分布频率示例

常用词元	词首分布频率	词中分布频率	词尾分布频率
系统	0.0041	0.0372	0.9587
器	0.0000	0.1617	0.8333
飞行	0.3273	0.3909	0.2909
机	0.1691	0.3873	0.4564

主题词词典中的词条的词性对于我们进行主题词识别也有较大的指导作用。通过对原始词表进行细粒度的词元拆分,确定词元最常用的词性。

此外,我们还可以分析词元是否是分割标记(如"、","," 等符号)、是否是数字、长度信息等构词特征。

2.2 统计信息特征抽取

词可以看做是字在满足一定情况下的组合,因此,本文提出的主题词抽取的任务也可以看做是从给定的语料中获取紧邻字组合成词的可能性,当这个可能性大于某个设定的条件,就认为可以将其作为一个完整的词条。我们用互信息作为度量这个可能性的大小。

互信息(mutual information)是用于描述两个变量之间关联程度大小的度量,其定义为

$$I(X,Y) = \log \frac{p(X,Y)}{p(X)p(Y)} \tag{3}$$

其中,$p(X,Y)$ 表示词元 X 与 Y 紧邻的情况下出现的概率,$p(X)$ 和 $p(Y)$ 是词元 X、Y 单独出现的概率。另外,从定义中可以看出,$I(X,Y) = I(Y,X)$,因此,实验中只需要计算一个词元与其前一个位置词元的互信息值。

为了使模型的训练速度更快,我们将连续的两个词元之间的互信息值进行离散化,主要是通过选取一个特定的阈值 θ,当两者的互信息值大于 θ 时,将值设置为 1,否则设置为 0。该阈值 θ 是通过对已有词库与文档计算出的互信息值,应选取合适的值。

2.3 维基百科的语法标记特点

维基百科采用 MediaWiki 作为其软件基础,具有自己的语义标记特点,形成了用户进行编辑 wiki 知识遵守的一种规范。

维基百科采用一个页面(Page)表示用户添加的一个词条,每个页面属于一个或者多个分类(category),每个页面内部还会有各种链接,包括内部链接(internal link)以及外部链接(external link),词条之间还会存在引用(citation)关系。这些关系通过 MediaWiki 的标记符号进行表示,对于标签的抽取具有很大的帮助作用。

表 3 中对 MediaWiki 的部分语义标记进行了说明。

我们利用 Lucene 提供的 WikipediaTokenizer 对从维基百科获得的文本进行分词处理,然后选取我们认为对构成主题词具有较大分辨作用的标记,并将其加入到词的特征向量中。

表 3 MediaWiki 语义标记及含义示例

语义标记	含义
'''X'''	表示 X 需要采用加粗字体显示
[[X]]	表示 X 是一个链接
Tag:X	表示 X 是一个标记,比如分类(Category)、模板(Template)、文件(File)等
=X=	表示 X 是一个内容区块,可嵌套,比如参考文献、注释等
*X	表示 X 是 list 的一个元素,可以嵌套
{{X}}	X 是一个需要特殊显示的区域,比如 about 等

另外,由于维基百科中的中文版包含有繁体、简体等多个版本,因此,需要在进行处理时考虑编码转

换的问题,以保证一致性。

3 实验以及结果分析

实验通过对比基于条件随机场的方法和单纯基于规则的方法。条件随机场采用CRF + + 进行模型的训练与测试,对比实验采用的是基于概率句法模式的识别方法 EPSyP[35]。

EPSyP 通过对构成实体名称的内部概率句法模式进行总结,得到一个泛化后的实体名称内部句法规则。这些规则全部是基于手工整理的。表 4 是部分使用的规则,字母均为词性标记,x + 表示有大于等于一个的词性 x,其中 a 为形容词,b 为区别词,n 为名词,g 为语素词。

表 4 EPSyP 句法规则举例

规则	准确率
a + b + n	1.000
a + g + n +	0.903
v + n +	0.850
g +	0.750

3.1 数据集

本文研究的领域是“太空”领域,使用的原始词表共有 2772 个词条,经过前后缀的抽取算法处理以后,共得到常用前缀 109 个,常用后缀 118 个。

本文选取的种子页面是分类“太空”,因此在 JWPL 中设置查询(Category:太空),然后使用广度遍历的方式对数据库进行检索,获取得到 2020 个页面;数据清洗阶段,去掉包括童话、小说、电影以及其它不相关的页面,最后剩余 1936 个页面;然后对这些文档进行转码,通过转码器将繁体中文转化为简体中文。

3.2 特征生成

根据第 2 节介绍的特征,经过特征抽取,得到采用的条件随机场的特征如表 5 所示。

在基于条件随机场的主题词识别方法中我们将每个词元标记为 B,I 和 O 三种标签,其中 B 表示该词元在主题词的开始部分出现,I 表示词元在主题词的其他部分出现,O 表示不是主题词的组成部分。

3.3 实验结果与分析

本实验中采用的衡量指标准确率 P、召回率 R 的定义如下:

$$P = \frac{N_s}{N_p} \tag{4}$$

表 5 条件随机场使用的特征以及含义

特征	含义
当前词元	如“载人”等词元,取值范围为所有词元
词性	如词性“b”,取值范围为所有词性
是否是常用前缀	取值范围为{IS _ PREFIX , NON _ PREFIX}
是否是常用后缀	取值范围为{IS _ SUFFIX, NON _ SUFFIX}
是否是断句标记	比如“、”、“,”等符号,取值范围为{IS _ SEG,NON _ SEG}
是否是数字	比如 202、731 等数字,取值范围为{IS _ NUM, NON _ SUM}
是否在标题中出现	该词元是否出现在该维基百科的页面的标题中,取值范围为{0,1}
是否为加粗显示	该词元是否是属于被 MediaWiki 加粗标记,取值范围为{0,1}
是否为一个类别	该词元是否属于维基百科的类别(Category),取值范围为{0,1}
是否为一个主要显示部分	该词元是否属于维基百科的主要部分(Heading),取值范围为{0,1}
是否是维基百科的内部链接	该词元是否属于维基百科的内部链接(Internal Link),取值范围为{0,1}
与前一个词元的互信息是否大于阈值 θ	与前一个词元的互信息是否大于阈值 θ,取值范围{0,1}

$$R = \frac{N_s}{N_r} \tag{5}$$

其中 N_s 为识别出来的正确的不同的主题词数目,N_p 为识别出来的所有的不同的主题词数目,N_r 为测试文档实际的不同的主题词数目。

为了更公平地衡量系统的指标,引入 F 值,解决可能存在的 P 和 R 相互矛盾的状况:

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 P + R} \tag{6}$$

其中 β 值用来调节 P 和 R 值的权重比例。在本实验中取值为 1。

我们对人工标注的结果集进行训练与测试,通过 10 折交叉验证的方式,然后计算最终的平均值,得到的结果如表 6 所示。

表 6 CRF 模型与 EPSyP 方法的实验结果比较

方法	P	R	F 值
CRF	0.941	0.753	0.836
EPSyP	0.743	0.767	0.755

从表6中可以看到,使用EPSyP方法比使用CRF具有稍高的召回率,这主要在于人工总结的大量的构词规则,但是,在准确率方面却明显不如CRF,这是因为大量的规则会使得结果中产生毫无意义的词汇,表7列举了部分识别错误的例子。

表7　EPSyP方法识别错误的主题词举例

所用规则	错误例子
n + v	会合周期是
vn +	处在赤道位置
vn +	称为地球静止轨道

从表7中可以看出,由于EPSyP方法只是从词性角度进行考虑,因此会在遇到反例情况时无法进行有效的排除。

CRF与EPSyP相比召回率稍有下降,但是准确率有较大的提升,使得整体的F值提高近8%。表8是采用CRF获得的主题词。

表8　CRF模型获得的主题词举例

火星静止轨道	载人航天计划
火星同步轨道	载人航天工程
卫星导航系统	航天发射场
卫星发射中心	航天发射基地
预定轨道	抛物线弹道
月球车	国际太空站
航天适应综合征	亚轨道飞行

从表8中可以看出,条件随机场模型对于长距离依赖以及多特征的统一建模具有很好的解决能力。

图2给出了使用这两种不同方法进行主题词抽取的接收者操作特征(ROC)曲线,由图可知使用CRF的识别主题词的能力优于EPSyP。

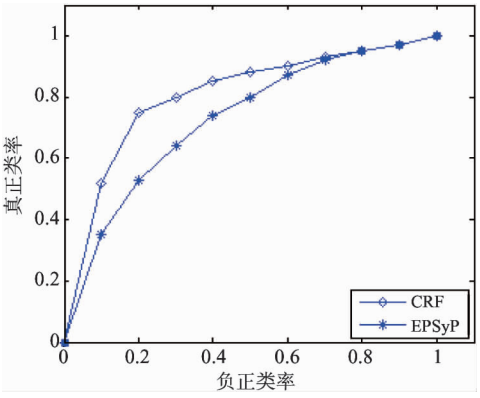

图2　主题词抽取使用不同方法的ROC曲线

在使用条件随机场的过程中,我们发现如下需要解决的问题:

(1)词元拆分过于散乱,造成词义丢失。

例如,“发射载具”一词,本意是指火箭、航天飞机等载体,然而在分词的过程中,由于词表容量的限制,无法收录“载具”这个词元,这就造成了将之拆分成“载”和“具”两个单字,其表征意义就明显下降。因此对于受限领域基础词元的工作需要进一步加强,以此来提高在分词的效果。

(2)对于命名实体识别能力不够。

例如,“西安卫星测控中心”、“酒泉卫星发射中心”这类词汇中,“西安”与“酒泉”均为实体名称,如果知道它们都是命名实体,就可以充分利用命名实体的特点来提高识别的效果。

对于上述问题,我们进行了初步的实验,实验结果表明,在给定可以完全覆盖住给定测试语料的基础词元和命名实体库的情况下,准确率可以提升3% ~4%。

在将来的工作中,我们将加强对基础词表的管理与更新工作,以发现较短且具有意义的词元,这对于发掘组合主题词具有很大的帮助作用;命名实体识别领域的进展对于本文的工作也是具有很大的促进作用,需要更深入的研究。

4　结　论

本文主要讨论了采用条件随机场进行领域主题词获取的方法,通过对维基百科的离线数据库进行数据收集,然后进行自动的特征抽取以及可视化标注,在此基础上,综合采用了基础词表的构词特点分析、词本身的特征、上下文互信息,以及对于维基百科的标注记法,将这些丰富的特征进行统一建模,取得了较好的实验效果。

为了提高主题词获取的效果,我们将针对基础词库的更新以及命名实体识别准确率的提高等问题继续进行进一步的研究。

参考文献

[1] Wikipedia. 维基百科:叙词表. http://zh.wikipedia.org/wiki/叙词表: Wikipedia, 2013
[2] Wikipedia. 维基百科官网. http://www.wikipedia.org/:Wikipeida, 2013.
[3] Chang J S, Lin Y C, Su K Y. Automatic construction of a Chinese electronic dictionary. In: Proceedings of the Third Workshop on Very Large Corpora. 1995. 107-120
[4] 仲云云, 侯汉清, 杜慧平. 电子政务主题词表自动构建研究. 中国图书馆学报, 2008, (003): 97-102
[5] 蒋维, 郝文宁, 杨晓恕. 军事训练领域核心本体的构建. 计算机工程, 2008, 34(5): 191-192

[6] Xu H, Yu B. Automatic thesaurus construction for spam filtering using revised back propagation neural network. *Expert Systems with Applications*, 2010, 37(1): 18-23
[7] Hearst M A. Automatic acquisition of hyponyms from large text corpora. *Association for Computational Linguistics*, 1992. 539-542
[8] Takenobu, Makoto, Hozumi. Automatic thesaurus construction based on grammatical relations. *International Joint Conference on Artificial Intelligence*, 1995,1308-1313
[9] Lassi M. Automatic thesaurus construction. Sweden: University Collage of Boras, 2002. 98-105
[10] Park Y C, Choi K S. Automatic thesaurus construction using Bayesian networks. *Information Processing & Management*, 1996, 32(5): 543-553
[11] Yang C C, Luk J. Automatic generation of English/Chinese thesaurus based on a parallel corpus in laws. *Journal of the American Society for Information Science and Technology*, 2003, 54(7): 671-682
[12] Hagiwara M Y, Ogawa, Toyama K. PLSI utilization for automatic thesaurus construction. *Natural Language Processing*, 2005, 334-345
[13] Tseng Y H. Automatic thesaurus generation for Chinese documents. *Journal of the American Society for Information Science and Technology*, 2002, 53(13): 1130-1138
[14] 王石, 曹存根. WNCT: 一种 WordNet 概念自动翻译方法. 中文信息学报, 2009, 23(004): 63-70
[15] 曾建勋, 常春, 吴雯娜等. 网络环境下新型汉语主题词表的构建. 中国图书馆学报, 2011,37(4):43-49
[16] 曾文. 网络化数字化时代主题词表自动构建技术的探索与实践. 国家图书馆学刊, 2012,(4):78-82
[17] 叶春蕾, 冷伏海. 技术路线图中未来技术词表构建方法研究. 现代图书情报技术, 2013,5: 59-63
[18] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, 2001. 282-289
[19] Sha Fei, Pereira Fernando. Shallow parsing with conditional random fields. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. 2003. 134-141
[20] Taskar, Abbeel, Koller. Discriminative probabilistic models for relational data. In: Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, 2002. 485-492
[21] Peng, McCallum. Accurate information extraction from research papers using conditional random fields. *Information processing and management*, 2006, 963-979
[22] 邓箴, 包宏. 基于条件随机场的中文自动文摘系统. 西安石油大学学报(自然科学版), 2009, (01): 96-99,102,114
[23] 张开旭, 夏云庆, 宇航. 基于条件随机场的古汉语自动断句与标点方法. 清华大学学报(自然科学版), 2009,(10): 1733-1736
[24] 郭家清. 基于条件随机场的命名实体识别研究:[硕士学位论文]. 沈阳:沈阳航空工业学院, 2007
[25] Sato K, Sakakibara Y. RNA secondary structural alignment with conditional random fields. *Bioinformatics*, 2005, 21(Suppl 2): ii237
[26] McDonald R, Pereira F. Identifying gene and protein mentions in text using conditional random fields. *BMC bioinformatics*, 2005, 6(Suppl 1): S6
[27] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. *Association for Computational Linguistics*, 2004, 104-107
[28] He X, Zemel R, Carreira. Multiscale conditional random fields for image labeling. *Computer Vision and Pattern Recognition*, 2004,(2):695-702
[29] Wang Y, Ji Q. A dynamic conditional random field model for object segmentation in image sequences. *Computer Vision and Pattern Recognition*, 2005, 1:264-270
[30] Kumar S, Hebert M. Discriminative fields for modeling spatial dependencies in natural images. *Advances in neural information processing systems*, 2004. 16(1-8): 29
[31] Zesch T, Müller C, Gurevych I. Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation, 2008. 1646-1652
[32] Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS. In: Proceedings of the second SIGHAN workshop on Chinese language processing, 2003. 184-187
[33] Apache Lucene. http://lucene.apache.org: Apache, 2013
[34] Kudo K. CRF++: Yet another CRF toolkit. http://crfpp.sourceforge.net, 2005
[35] 王石. 中文实体名称的识别和语义分析方法研究:[博士学位论文]. 北京:中国科学院计算技术研究所, 2009

A method for domain-specific subject word extraction based on Wikipedia and conditional random fields

Qi Baoyuan[*,**], Shi Zhongzhi[*]
([*] Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)
([**] University of Chinese Academy of Sciences, Beijing 100190)

Abstract

Aiming at the shortcomings of time-consuming and long update cycle of the traditional manual thesaurus creation, a new method for acquiring subject words for creating or updating a specific field's thesaurus based on the Wikipedia and conditional random fields (CRF) was presented. The method fully uses the Wikipedia's unique structural advantage and co-edited encyclopedic knowledge in various fields to automatically obtain new subject words according to the existing thesaurus's characteristics in lexical structure and statistical distribution to replenish the existing thesaurus. The method trains a CRF model to acquire new subjects from Wikipedia, and synthetically models various features to achieve the better experimental effect. The experimental results show that the method can increase the *F*-value of subject identification up to about 83%.

Key words: thesaurus construction, subject word extraction, Wikipedia, conditional random fields (CRF)