基于百科词典的知识获取系统的研究与实现①

许勇[1] 宋柔[2]

[1](北京工业大学计算机学院 北京 100022)

[2](北京语言文化大学计算机系 北京 100083)

E-mail: hopexy163@163.com

摘要： 从各种自然语言文本中获取知识是自然语言处理技术的重要应用。本文描述了从百科辞典文本中获取知识的探索性的研究工作，介绍了一个实验性的，限定范围的百科辞典知识获取系统。具体工作包括：利用分词工具进行初步的词条分类；在词条分类的基础上，对处理范围内的词条文本进行观察，以人工方式归纳其中目标知识的基于语义特征的模式规则；利用 YACC 工具对模式规则进行解释，进而抽取目标知识。文中给出了试验结果及分析。

关键词： 自然语言处理，知识获取，信息提取

1. 引言：

利用计算机从各类自然语言文本中获取知识是自然语言处理技术的一个应用领域，它可以满足未来计算机发展的智能化与知识化需求的一方面。目前来讲，从未加限制的开放文本中获取知识是不太现实的。一种非常自然的方式是将体裁限制于词典的文本。这是因为：

1. 词典文本中的知识比较密集，从词典中获取知识效率较高。
2. 词典文本的表达方式比较有规律，与计算机的处理能力较为相匹配。

词典文本知识获取的工作在国内外都在展开。有两种类型的工作。一是面向计算机的，用以支持自然语言处理应用系统的开发。Princeton University 的 WordNet 是利用词典文本开发的庞大的英语词汇知识库，我国董振东教授的 HowNet 是有相当规模的中文词汇语义知识库，它们都是手工开发的。另一类工作是面向人的。微软公司的电子百科全书 Encarta 能提供词语检索、领域检索、媒体类型检索、时间检索、地理区域检索等检索手段，帮助人从这个词典中获取百科知识。但是它只能以词条文本全文的形式提供知识。

近年来兴起了一种和词典文本的知识获取关系非常密切的技术——信息提取（Information Extraction）。知识也是一种信息，如果把要提取的知识看成感兴趣的信息，那么，这个技术就可以应用于词典文本中的知识自动提取中。信息提取技术的应用目的是在某一领域的文档集合中自动抽取指定信息，加快人们获取、检索信息的速度。信息提取系统一般面向特定的领域开发，在应用领域发生改变的时候一般都需要一定程度的重新构建。这项技术中比较重要的是用于描述目标信息的句法或者语义结构的模式规则，系统根据这些模式来识别、提取目标信息。获取模式的方法主要有人工获取和机器自动获取等方式，在后一种方式中大量采用了机器学习方法。人工方式的性能高，但开发量比较大。自动学习

①本论文得到了以下项目和基金的支持：教育部科学技术研究重点项目（编号 00128）

方式能节省大量人力，但目前来说其性能还有待提高。

2. 面向百科词典的人机结合知识提取系统ENCYLIB的设计实现

文本知识提取和信息提取在技术、方法上很接近。但知识提取还是有和信息提取相区别的一些特点。知识是信息的一种，它比一般信息更有普遍性，有更高的稳定性，也更具系统性。另外一点是，一般信息提取系统是针对某个特定的目标领域来开发。而知识提取可以是领域限定的，也可以不是。知识提取可以对几个相互关联的领域的文本进行处理，构筑基于这几个领域的内部互相关联的知识库系统。

我们选择了《中国大百科全书》的电子版作为知识提取试验系统的对象。这个词典的文本具有体裁一致，题材多样等特点，并且不少卷目的文本中存在比较统一的信息表述模式，这对计算机的处理是比较有利的。

2．1 系统的功能设计

目前的处理范围包括以下几个部分：

1. 《中国地理》卷目中的行政地名词条。

抽取的项目：行政隶属关系；行政功能；面积；人口；地理方位；经纬度。

其中，行政隶属关系指的是上级行政地名的列表。行政功能指的是是否省会、首府、地区行署驻地等等。

2. 《美术》卷目中的外国美术家人名词条。

抽取的项目：国籍；职业；性别；出生年月日；出生地；死亡年月日；死亡地。

之所以选择人名和地名这两种题材的条目，是出于以下两点考虑：一是普遍性较高。人名和地名是比较重要的两类专名，在百科词典的不同卷目中大都会包含一定量的人名与地名。二是这两类词条的文本在信息表述形式上较具一致性。这是指这样一种情况，某些相同类型的信息在不同的文本中以相近的位置、相近的句型重复出现。其它类型的词条文本之间的一致性没有这两种词条来的好，大多过于“散乱”。

2．2 实现方法

从信息提取系统的角度看，本系统目前采用的方法是人机结合方式。机器进行条目自动分类，并使用模式在词条文本中进行知识提取；模式规则是靠人工观察获取的。在选定了要处理的卷目与词条类别范围之后，我们观察了需要处理的词条文本，确定了要提取的知识（信息）的范围，然后用人工方法归纳了相关的信息出现的规则。

从研究的角度看，机器自动获取模式的方法显然是更吸引人的，而且这种方法也是今后的改进方向。目前阶段采用人工方法是遵循人的认识规律，从特殊到一般地认识事物。人类专家在人工归纳模式的过程中，摸索一般规律，为指导机器自动提取模式打下基础。

2．3 系统的处理流程和模块构成

系统目前主要以下有三个部分构成：

1. 词条的自动分类。

2. 《美术》卷目的外国美术家人名词条，《中国地理》卷目的行政地名条目中目标知识的抽取处

理。

3. 查询以上两类知识的查询模块。

词条的自动分类主要是借助于分词系统，针对人名、地名、机构名条目的识别。目前查询模块比较简单，把抽取的结果转化为数据库，实现了简单的查询界面。下面主要介绍第二个部分。

2．4 人物、地名条目中知识获取

要实现自动抽取，需要做以下三个部分的工作：1．归纳模式规则，收集特征词汇。2．词条文本的分词，并对特征词汇做出标注。3．对文本进行分析，抽取相关知识项目。

这里指的特征词汇是模式规则中牵涉到的具体的词汇项。第二个问题可以借助于分词系统解决。我们实验室的通用分词系统可以挂接用户词库，能较好地满足不同应用对于分词的不同颗粒度的要求和处理新词的要求。归纳模式规则，则涉及到几个问题：

1．应该基于自然语言的什么特征。2．规则的形式与其解释。

本系统中的模式规则是基于语义特征的。之所以采用基于语义特征的模式，是因为知识与语义特征有直接联系，与句法特征并无直接联系，而且至今没有获得实际应用效果好的句法分析系统。语义特征虽然在没有限制的、开放的范围内难以形式化，但是在某种方式限定的范围内可以比较好的归纳其规律。

鉴于上下文无关文法的成熟性和易用性，在本系统的规则中描述句子结构的部分采用了单纯的上下文无关文法形式。采用单纯的上下文无关文法形式的模式规则，其解释可以利用现有的软件工具，如YACC。这种工具可以生成指定文法的解释程序，可以大大方便规模不是很大的文法的解释程序的开发。在本系统的实现中，就使用了YACC的一个版本——BYACC。

YACC的输入是一个说明文件，输出是一个由分析程序的C源代码组成的代码文件。YACC中除了可以指定文法规则之外，还可以直接书写可供分析过程中调用的C程序代码。这些代码称之为“动作”。利用YACC的文法规则指定相关信息的模式规则，利用YACC中的动作来实现抽取操作。这样，在本系统中，模式规则是YACC规则，特征词汇是YACC规则中的终结符号，具体抽取操作由YACC动作来完成。下面选择介绍两个卷目中的模式。

2．5 《美术》卷目中西方美术家词条中相关信息的模式规则

在这个卷目的大部分词条文本中共同出现的知识项目是美术家的国籍、职业、性别、出生时间、出生地点、死亡时间、死亡地点共7个项目。这7个项目在句子中的分布情况是这样的：国籍、职业、性别三个项目出现在一个句子中，出生时间、出生地点出现在一个句子中，死亡时间、死亡地点出现在一个句子中。需要说明的是，这里指的句子并不包括复句形式的句子。因此，句子的结束标点不是句号，而是逗号、分号和句号。这样限定的理由是因为对复杂的复句形式的句子的结构规律做出归纳是非常困难的。

由于目标项目主要分布于三个句子，从归纳模式规则的角度看，就是给这三个含有目标信息项目的句子的结构模式做出归纳。因此，这个卷目中的最上层的规则如下：

phs : gujzhy | chushd | siwngd ;

这是YACC形式的规则，可以将这个规则写成等价的BNF范式如下：

phs→gujzhy | chushd | siwngd

phs是这个卷目中规则的总的开始符号。gujzhy，chushd，swngd依次是代表国籍职业句，出生地、出生时间句，死亡地、死亡时间句的文法符号。gujzhy代表的国籍职业句主要由三个内容成分组成。按照顺序分别是国籍部分、时代部分、职业部分。这三个成分以如下两种模式构成国籍职业句：

1. 国籍+时代+职业

例：意大利 文艺复兴时代 画家、科学家。

国籍 时代 职业

2. 国籍+职业

例：西班牙 画家。
国籍 职业

根据以上规律，这个句子的第一层规则如下：

gujzhy ： guoji shidai zhiyel | guoji zhiyel ;

上式中，guoji 代表国籍成分；shidai 代表时代成分；zhiyel 代表职业成分。

这三个内容成分各自有自身的结构。国籍成分主要由表示民族血统的族裔成分、国名成分组成。这部分的规则如下：

guoji ： zuyi guoj | guoj | zuyi ;

zuyi ： GUOMJ YIJI | GUOM YIJI;

guoj ： GUOM | GUOM GUOM | GUOM DANGANG GUOM ;

上式中，zuyi 代表族裔成分，guoj 代表国名成分。国籍成分可以以族裔后跟国名的形式构成，也可以单独由族裔，或者国名的形式出现。例如，“美籍德国画家”，“德国籍画家”或者“德国画家”等。最后一种形式是最常见的。族裔部分有两个子部分，国名简称和“籍”或“裔”等词汇。上式中，GUOMJ 代表国名简称，GUOM 代表国名。YIJI 代表“籍”类的词汇。国名部分除了简单的由国名组成之外，还有两个国名重叠，或两个国名之间由“-”号连接的形式。比如“南斯拉夫克罗地亚画家”，“西班牙-法国画家”等。上式中，大写的符号是终结符号，在规则中是最下层的符号，本身没有结构。这些终结符号就是特征词汇，大部分情况下，对应于汉语中的一个词。其他非终结符号都有自己的下层结构，直到特征词汇。具体的抽取相关知识项目是通过 YACC 的动作来完成的。下面以国籍项目为例来说明利用 YACC 动作来抽取的方法。

guoj ： GUOM{$$ = $1;} | GUOM GUOM{$$ = $1+$2;} | GUOM DANGANG GUOM{$$ = $1+$2+$3;} ;

这个规则上面已经说明过。与上面不同的是，规则的每个选择项后面多了动作。YACC 动作的调用发生在动作依附的规则被识别出来之后。调用、赋值、维护等操作都由 YACC 内部机制来完成。$$，$1，$2，$3 等是 YACC 伪变量，分别代表规则左部符号对应的值和规则右部的第 1，2，3 个符号对应的值。直观的讲，如果第二个规则识别成功，调用动作{$$ = $1+$2;}，之后与 guoj 符号联系的值就会是原句子中与 GUOM GUOM 这个模式相匹配的连续两个国名字符串。第一，第三个规则也是一样。对具体句子“西班牙-法国画家”中的“西班牙-法国”句子片断来说，上面的规则会匹配第三个选择项，并成功规约出非终结符 guoj。经过相应的动作调用之后，与 guoj 联系的值就会变成字符串“西班牙-法国”。其他抽取项目的抽取方法雷同。这些抽取项目再经过验证，就可确认为对应条目的知识项目。

2．6《中国地理》卷目中行政地名词条中相关信息的模式规则

这个卷目中的行政地名词条中，共同出的信息项目是行政隶属（上位行政地名）、行政功能、地理方位、面积、人口、经度纬度共 6 个。这 6 项信息在句子中的分布情况是这样的：行政隶属、行政功能共同出现在一个句子中，其余 4 个项目各自出现在不同的句子中。因此，需要给 5 种不同的句子归纳规则。为了下文中说明方便，出现行政隶属、行政功能的句子简称为行政隶属功能句。其余分别称为地理方位句，面积句等。这 5 种句子中，行政隶属功能句的模式比较复杂，主要分成三种类型的结构。下面是这个卷目规则的最上层规则。

phs ：cgxz | cgjt | ssmz | wydlwz | cgmr | jingwei ;

上式中，phs 是文法的开始符号。cgxz,cgjt,ssmz 分别描述行政隶属功能句的三种不同类型。wydlwz 描述地理方位句的结构，cgmr 描述面积、人口两个句子的结构。jingwei 描述经纬句的结构。行政隶属功能句主要有三种内容类型。依次是单纯行政隶属方面的内容，经济特征方面的内容，所居住少数民族方面内容。以下是这三种类型的例子：

“阿克苏地区辖市和行署驻地”；“河北省轻工业城市”；“吉林省朝鲜族主要聚居地”

以上三种类型的句子在规则中分别对应于 cgxz,cgjt,ssmz 三个文法符号，第一种类型的句子的规则是这样的：

cgxz : sxzd XIA xzd | sxzd XIA xzd LIAN xzgn | sxzd xzgn | sxzd xzd
| sxzd xzd LIAN xzgn | XIA sxzd ;

上式中，sxzd 是上级行政单位部分的文法符号，XIA 是“辖”、“属”等表示行政隶属关系的特征词汇，xzd 是行政单位名，如“省”、“市”、“县”等。xzgn 表示行政功能，如上例中“行署驻地”就是行政功能。LIAN 是起到连接作用的特征词汇，如“和”、“与”等。上级行政地名部分内部结构是地名序列，大部分情况下都是行政性地名，但也有少数的句子中包含自然地理地名。其他类型的句子和对应的非终结符号也都有各自的下层结构。具体的抽取知识的方法和《美术》卷目中的方法一样。

表 1 两个卷目中规则的情况

卷目	规则数	特征词类个数	特征词数
《美术》西方美术家人名	76	31	654
《中国地理》行政地名	152	61	555

3. 实验结果和分析

两个卷目的抽取情况如下表：

表 2《中国地理》卷目行政地名实验结果

	遗漏数	错误数	正确抽取数	有相关信息的词条数	召回率（%）	正确率（%）
行政隶属	60	12	646	716	91	98
行政功能	5	0	121	126	96	100
面积	2	1	710	712	99	99
人口	3	0	723	726	99	100
地理方位	25	2	721	746	96	99
经纬	1	1	28	30	99	99
词条总数	746					

表 3《美术》卷目行政地名实验结果

	遗漏数	错误数	正确抽取数	有相关信息的词条数	召回率（%）	正确率（%）
国籍	3	0	425	428	99	100
职业	3	0	426	429	99	100
出生地	6	0	400	406	98	100
出生时间	0	2	427	429	100	99
死亡地	6	0	327	333	98	100
死亡时间	0	4	398	402	100	99
词条总数	429					

这两个卷目的抽取失败中遗漏的情况占大多数。收录到系统模式规则库的模式需要有一定的一般性，也就是说一条规则覆盖的实例要尽量的多。如果对每个出现次数较少的“特殊”实例都写“特殊”规则的话，规则的规模就会变得很大，系统就会变得复杂，难以进行维护。因此，对一些出现次数较少的特殊形式的实例采取了放弃的方法。这样一来，就难免不能识别一些生僻的实例。下面是一个抽取遗漏的

例子。

台北县

台湾省人口密度最高的县。

在这个例子中，修饰地名通名“县”的部分是一个独立的短语结构的“人口密度最高”。以独立的短语结构形式构成修饰部分，其结构、内容可以是千变万化的，很难用适量的规则来概括描述。抽取错误的情况主要是规则库中的一些规则覆盖了不属于含有目标知识的句子所致。如：

海宁市

行政隶属: 浙江省,嘉兴市

这是一个抽取错误的例子。此例的条目是海宁市，抽出的行政隶属项目中其上级行政隶属单位有浙江省、嘉兴市两个行政单位。其中，浙江省是对的，嘉兴市就错了，海宁市不属于嘉兴市。之所以造成这个错误是因为在词条文本中有如下叙述：

海宁市

浙江省蚕茧、络麻和油菜籽重点产区之一。

。。。 。。。1986 年撤县设市，属嘉兴市管辖，

1988 年改为省直辖行政单位。 。。。 。。。

从以上说明中可以看出，海宁市在 1986 年到 1988 年期间由嘉兴市管辖，到 1988 年以后改为省直辖行政单位。文中“属嘉兴市管辖”一句是造成错误的原因，因为在规则库中有对应于这类句子的模式规则。针对这个错误，可以有几种特殊的方法解决，但是从信息抽取的角度看，这种错误需要做超出句子范围的段落一级的处理才能解决，是具有一定普遍性的错误。《美术》卷目抽取失败的情况大致上差不多，但是《美术》卷目中含有目标知识的句子比《中国地理》卷目的文本的句子相对简单一些，还有处理包括的词条数要比《中国地理》少，因此性能指标高了些。

4. 今后的工作

目前本系统尚处于实验性的阶段，很多方面有待改善、扩充。今后的研究主要包括抽取方法的改进，处理范围的扩展和对抽取结果的进一步改进等。抽取方法的改进方向主要是引入机器学习方法，提高系统的自动化水平。目前系统只含有两个卷目两个类别的知识。作为一个基于百科全书的知识库系统而言，这个范围显然是不够的，有必要进行扩展。

参考文献

[1]许勇，宋柔，基于百科辞典的知识获取系统的研究与实现，北京工业大学硕士论文，2001

[2]J.E.霍普克罗夫特，J.D.厄尔曼，自动机理论、语言和计算机导引，科学出版社，1986

[3] Stephen G. Soderland,CRYSTAL:Learning Domain-specific Text Analysis Rules，1996

[4]Clairec Cardie,Domain-Specific knowledge Acqusition For Conceptual Sentence Analysis,1994

作者简介： 许勇,男,吉林人,博士研究生,研究方向为自然语言处理;宋柔,男,江苏苏州人,,教授,博士生导师,主要研究领域为自然语言处理，人工智能.

An Experimental Encyclopedia-Based Text Knowledge Acquisition System

Xuyong[1] Songrou[2]

[1] (*Beijing Polytechnic University, Beijing 100022,China*)

[2] (*Beijing Language and Culture University, Beijing 100083,China*)

E-mail: hopexy163@163.com

Abstract: This paper presents an experimental system of acquiring knowledge from encyclopedia-text. The source encyclopedia is "*Encyclopedia of china*", and currently the system includes two kind of items: Chinese district item from "*China Geography*" volume, foreign artist item from "*Art*" volume. The system consisted of three main modules: Encyclopedia-item classification module, item-text analysis and knowledge-extraction module, query module. The domain extraction rules are semantic-feature based, and these were acquired by hand. YACC tool was adopted to analyze the item text. In the last, the testing result and it's analysis were presented.

Keyword: Natural Language Processing, Knowledge Acquisition, Information Extraction