一个面向辞典的知识获取系统的设计与实现

刘亮亮,李志庆,孙 颖
(江苏科技大学,江苏 镇江 212003)

摘 要:在知识获取中,手工填槽是一件繁琐而又枯燥的工作,效率很低。针对一类词条的处理提出了一个面向辞典的知识获取系统,通过分析辞典中文本的特征,最大可能地实现了填槽的机器自动生成,该系统分为三个子系统:词语识别子系统、规则匹配子系统、运行填槽子系统,完成了文本识别、规则匹配和运行填槽功能。

关键词:知识获取;槽;词语识别;产生式规则
中图分类号:TP182 **文献标识码**:A **文章编号**:1673-629X(2007)09-0099-03

Design and Implementation of a Dictionary-Oriented Knowledge Acquiring System

LIU Liang-liang, LI Zhi-qing, SUN Ying
(Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: In the knowledge acquiring, it is not only fussy and boring to fill slots manually, but also inefficient. Being faced with the manipulation problems of a type of vocabulary entry, presents a knowledge acquiring system which is dictionary-oriented. This system implements the automatic filling-slot through analyzing characters of texts in the dictionary. It can be divided into three sub-systems, that is words recognition, rules matching and slots filling, which can recognize the words, match the rules, and fill the slots.

Key words: knowledge acquiring; slot; words recognition; production rule

0 引 言

知识获取,是指知识从外部知识源到计算机内部的转换过程,就是从知识丰富的源头收集、整理、形式化知识,加入到知识库中。知识获取是建造知识库的一个瓶颈,也是知识工程领域研究的热点[1]。

辞典具有蕴涵丰富、密集的领域知识,语言表达较规范的特征,是一类十分可取的用于构建知识库的知识源。但相比结构化的文本而言,从辞典中获取知识仍然较为复杂。目前,从辞典中获取知识仍大多采用手工方式,费时费力,而且容易引入人为错误。文中针对一类词条(官名词条)的处理设计了一个面向辞典的知识获取系统,并讨论了该系统推广应用的必要条件。

1 系统设计的背景简介

1.1 系统设计的目的

系统设计目的:在最少、最简单的人工干预下,实现知识的框架表示中槽值的填写,即将以下词条:

三等参谋官 官名。清末新陆军军官。
光绪三十年(1904)定新陆军营制,始置。
每镇一人,正五品,正军校充,奏补。
随同正参谋官分任计划诸务。
——摘自《中国历史大辞典》

转化为框架表示形式[2]:

框架名:〈三等参谋官〉
是一个:官名
时期:清末
是一种:新陆军军官
设置时间:光绪三十年(1904)
设置原因:定新陆军营制
任职人数:每镇一人
官阶:正五品
任职人员:正军校
授官方式:奏补
职掌:随同正参谋官分任计划诸务

1.2 手工填槽方式简介

手工填槽按照如下的步骤进行:

(1)由知识工程师在领域专家的指导下给出知识

收稿日期:2006-12-07
基金项目:国家自然科学基金资助项目(60310213)
作者简介:刘亮亮(1979-),男,江西萍乡人,硕士研究生,研究方向为知识工程与信息处理;导师:张再跃,教授,研究方向为智能信息处理。

的框架描述,即给出槽名、侧面名,并对其进行注释。

(2)知识工作者在框架描述指导下进行具体填槽。

手工填槽是一件繁琐而又枯燥的工作。一方面框架中属性和关系众多,不容易全部熟悉,经常要先到框架描述中查询属性或关系的名称,再进行填写,效率很低。另一方面词典中的词条中有许多相同或相似的表达,很多槽值的填写只是简单的重复,工作枯燥[3]。因此,通过分析辞典中文本的特征,最大可能地实现机器生成,是一个需要尽快解决的问题。

2 面向辞典的知识获取系统的设计

面向辞典的知识获取系统有三个子系统,分别完成三项功能:

(1)对文本进行词语识别,为已识别的词加上类别,并将无法识别的词语输出。

(2)对完成词语识别后的文本,根据它们的类别组合寻找规则,将无法匹配的类别组合输出。

(3)对完成词语识别与规则匹配后的文本运行填槽,得出最终结果。

系统的功能结构主要包括词语识别子系统、规则匹配子系统和运行填槽子系统。

2.1 词语识别子系统

词语识别子系统根据分类词汇词典对文本进行自动词语识别,找出无法识别的词语。

2.1.1 分类词库

分类词库为自动词语识别提供语义支持。它的每个词条由词语和该词语的类别两部分组成。而这些词条大致可以分为三类:

(1)填槽用词:这一类主要是一些名词,例如:官名、机构名、朝代名、年号、官阶、俸禄等。

(2)槽名设置用词:这一类主要是动词和少数的名词,例如:置、沿置、罢、复名、掌、管理、长官、位等,它们的类别采用在前面加#号的方法处理。例如:#置、#沿置、#罢、#复名、#掌、#管理、#长官、#位等。

(3)词语切分用词:这类词对填槽以及设置槽名都没有帮助,仅在词语切分时有意义。标点符号可以看成一种特殊的该类词。其类别采用在前面加 & 号的方法处理。这类词不能轻易设置,关于这类词的判别有待进一步探讨。

根据实际情况,约定每一个词只有一个类别。这样可以简化词语的识别,并且基本没有产生问题。如果发生一个词有两个以上类别,例如既属于官名又属于机构名,这种冲突的情况也有待进一步探讨。

需要说明的是,大多数词语是通过其他手段大批找出的。例如与官名词条相关的几类词汇官名、机构名和年号等可以从中国历史大辞典中用计算机很快找出。因此,人工添入的词相对是少量的,这正是知识自动获取可行的一个重要条件。

2.1.2 词语识别的流程

词语识别子系统读入源文本(辞典),经过人机交互——主要是计算机自动识别,得到合理的词语识别结果。词语识别的流程如图 1 所示。

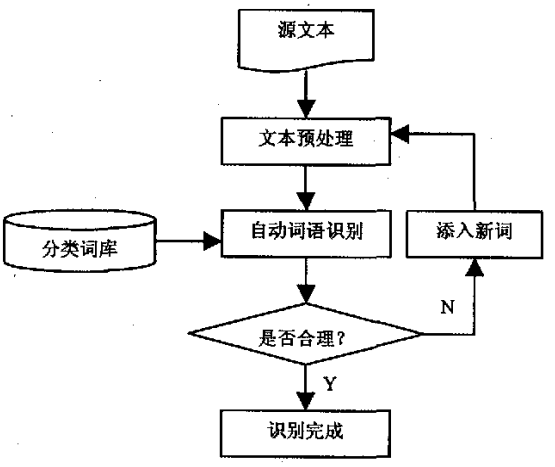

图 1 词语识别流程图

2.1.3 自动词语识别算法(伪代码描述)

```
自动词语识别
{
    读入预处理(预处理:文本错误修改、分段与分句等)后的文本;
    读入一段(循环);
    读入该段的一行(循环);
    对该行采用最大字串匹配法[4]正向逐词匹配:
      匹配成功,加上类别;
      匹配不成功,标识该字;
    对执掌类型词语进行处理;
    对相邻无法匹配字组合后输出,并同时输出该句;
}
```

该算法的关键在于字串的匹配,这关系到自动词语识别[4,5]的成败。

2.2 规则匹配子系统

规则匹配子系统根据词语类别的组合进行规则匹配,找出相应的规则与填槽时使用的槽名。

2.2.1 产生式规则的使用

产生式规则[2]是一个以"如果满足这个条件,就得出某一结论"形式表示的语句,其基本形式为:

IF 前提 THEN 结论

具体的,在规则匹配子系统中,规则的前提是词语类型的组合,而规则的结论是相应的槽名。相应的形式为:

结论:前提

例如:

俸禄: 俸禄
为官员之副贰: 官名 #副贰
沿置时代: 时代 #沿置
时代上级官员: 隶 官名
为机构次官: #为 机构名 #次官
时代为机构次官: 时代 机构名 #置 #为 #次官
时代所属机构: 时代 #隶 机构名

2.2.2 规则匹配的流程

规则匹配的流程如图 2 所示。

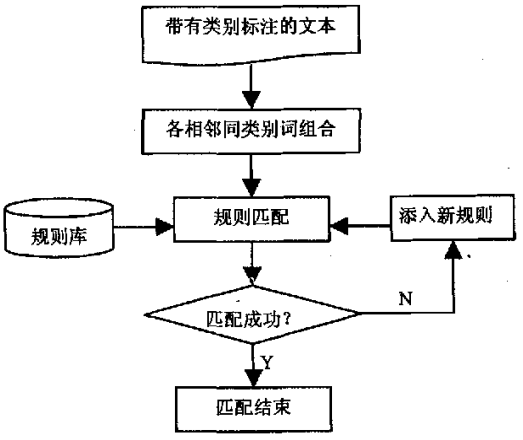

图 2　规则匹配流程图

规则匹配的算法就是根据前提寻找相应产生式的算法,此处不再赘述。

2.3 运行填槽子系统

文本经过前两个子系统的运行后,添入了新词与新规则,于是可以顺利地填槽了。运行填槽子系统使用新词库进行分词,使用新的规则库进行规则匹配(槽名的查找),在找到槽名后,将相应的填槽用词填入槽中。其流程如图 3 所示。

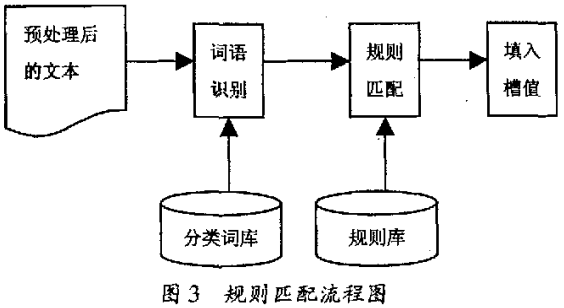

图 3　规则匹配流程图

3 面向辞典的知识获取系统的实现

笔者利用 Microsoft Visual Studio .NET 2003 开发环境,实现了面向辞典的知识获取系统,程序运行效果良好。笔者设计出了程序的三个功能界面,并逐步实现了相应功能。界面分别如图 4、图 5、图 6 所示。

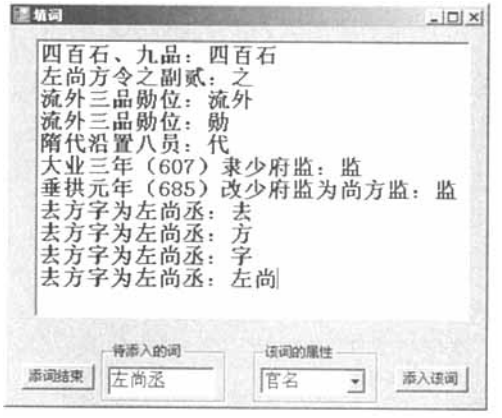

图 4　词语识别子系统的添词界面

图 5　规则匹配子系统的规则添入界面

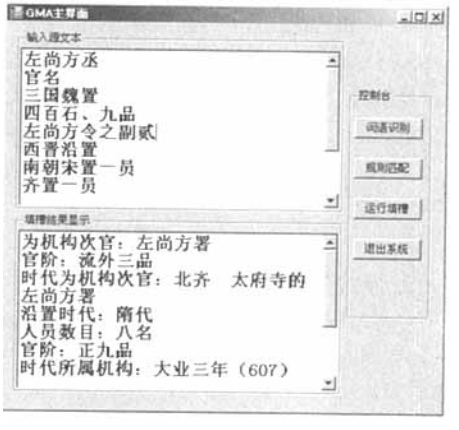

图 6　运行填槽界面(主界面)

4 结论与展望

本系统相比手工填槽有很大的改进,提高了效率,

(下转第 105 页)

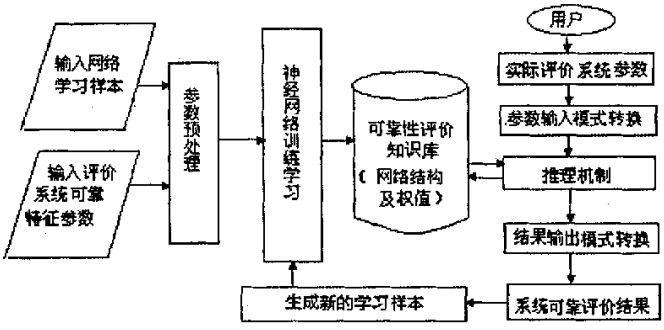

图 5　优化 BP 神经网络可靠性预测模型

度 R 作为输出变量。根据上述的可靠参数处理以及对 BP 神经网络的要求,将表 1 的数据转换为对 BP 神经网络可靠性预计模型的训练样本,训练达到一定精度后停止训练。当学习因子 $\eta = 0.9$,动量因子 $a = 0.6$,预设误差为 0.00001,单隐层,其隐节点数为 $L = 6$,所得到的网络评价的结果见表 1。

表 1　设备可靠评价学习样本

网络学习样本				
评价参数	设备 1	设备 2	设备 3	设备 4
1.资金投入	-1.5	0.5	-0.5	0.5
2.危险源状况	-0.5	-1.5	0.5	1.5
3.工作时间	-0.5	0.5	0.5	-0.5
4.故障概率	-1.5	-0.5	-0.5	0.5
5.防灾能力	0.5	0.5	1.5	-1.5
6.安全记录	-0.5	-0.5	0.5	1.5
评价结果 R	-0.5	0.5	0.5	1.5
	可接受	好	好	非常好

2)评价结果。

为检验网络训练的效果,将表 2 中的另 4 个设备的同样 6 个可靠参数数据输入,得到表 2。

表 2　设备神经网络可靠性评价结果

网络可靠评价结果				
设备	设备 1	设备 2	设备 3	设备 4
1	-0.5	-1.5	0.5	1.5
2	-1.5	0.5	-0.5	0.5
3	-0.5	0.5	0.5	-0.5
4	-1.5	-0.5	-0.5	0.5
5	0.5	0.5	1.5	-1.5
6	-0.5	-0.5	0.5	1.5
结论	-0.5	0.5	0.5	1.5
	可接受	好	好	非常好

表中结果表明基于优化 BP 神经网络的系统可靠评价模型的可行性。

4　结束语

BP 神经网络理论应用于系统可靠评价中,可利用神经网络并行结构和并行处理的特征,通过适当选择评价项目,能克服可靠评价的片面性,可以全面评价系统的可靠状况和多因数共同作用下的可靠状态。运用神经网络知识存储和自适应特征,通过适应补充学习样本,实现历史经验与新知识完满结合,在发展过程中动态地评价系统的可靠状态。且利用神经网络理论的容错特征,通过选取适当的作用函数和数据结构,可以处理各种非数值性指标,实现对系统可靠状态的模糊评价。文中将优化后的 BP 神经网络应用于系统可靠评价中,人工神经网络具有自组织、自学习和非线性逼近的能力,将神经网络运用于函数逼近,应用十分广泛,为系统的可靠性预计提供了一种新的方法。

参考文献:

[1] 蒋宗礼. 人工神经网络导论[M]. 北京:高等教育出版社,2001.

[2] 田景文,高美娟. 人工神经网络算法研究及应用[M]. 北京:北京理工大学出版社,2006.

[3] 杜　黎,陈　陶. BP 网络预测能力仿真与分析[J]. 昆明理工大学学报,2003,28(5):78－80

[4] Widrow B, Nguyen D. Improving the learning speed of 2－layer neural networks by choosing intial values of the adaptive weights[C]//In Proceedings of the IEEE International Conference on Networks. Amsterdam, The Netherlands: Elsevier Science Publishers B. V., 2000:21－26.

[5] 丛　爽,赵　何. 反向转播网络的不足与改进[J]. 自动化博览,1999(1):25－26.

[6] 楼顺天,施　阳. 基于 MATLAB 的系统分析与设计——神经网络[M]. 西安:西安电子科技大学出版社,1998.

[7] Relex Software Co. & Intellect. 可靠性实用指南[M]. 北京:北京航空航天大学出版社,2005.

(上接第 101 页)

在知识获取中有一定的使用价值。由于受计算机自动识别的能力以及规则匹配算法的限制,本系统还难以达到非常理想的效果,所以对这方面的改进将是笔者下一步的工作。

参考文献:

[1] 王文杰,叶世伟. 人工智能原理与应用[M]. 北京:人民邮电出版社,2004.

[2] 蔡自兴,徐光祐. 人工智能及其应用[M]. 第 3 版. 北京:清华大学出版社,2004.

[3] 徐　波,孙茂松,靳光瑾. 中文信息处理若干重要问题[M]. 北京:科学出版社,2003.

[4] 曹　倩,丁　艳,王　超,等. 汉语自动分词研究及其在信息检索中的应用[J]. 计算机应用研究,2003(5):71－74.

[5] 沈达阳. 汉语分词系统中的信息集成和最佳路径搜索方法[J]. 中文信息学报,1998,11(2):34－47.