《中国大百科全书》人物传记知识提取加工规范[1]

颜伟 王洁 尚英 宋柔

北京语言大学 语言信息处理研究所 北京 100083
E-mail: {yanwei, wangjie,shangying,songrou}@blcu.edu.cn

摘要： 将百科辞典中的知识形式化，是使用计算机进行知识服务的根本基础。目前的主要方法是在人工建立语言知识库的基础上计算机对于词条释文进行句法语义分析，或者直接由知识专家对词条释文进行形式化改写。这两种方法都需要大量的高级人力的投入。我们试图用计算机自动提取百科辞典中的知识，主要思路是降低要求，只提取有充分的、明确的形式特征的显性知识，并且先由人工模仿计算机进行提取，建立起显性知识的数据库，供计算机系统进行训练和测试。本文详细介绍了百科辞典中人物传记条目释文中显性知识的表示规范，包括知识点的取舍，简单知识和简单知识组的形式化表示方法，以及显性的复杂知识的提取方法。

关键词： 人物传记 知识提取 加工规范

The Specification for Biographic Knowledge Acquisition from Electronic Encyclopedia of China

Yan Wei Wang Jie Shang Ying Song Rou
Language Information Processing Center
Beijing Language and Culture University, Beijing 100083
E-mail: {yanwei,wangjie,shangying,songrou}@blcu.edu.cn

Abstract: The formalization of cyclopedia knowledge is the base of knowledge service by computers. This can be achieved either by computers analyzing the syntactic or semantic structures of the texts, or by knowledge experts rewriting the contents of the lists. Both rely on the knowledge data acquired manually, which requires a high input of advanced manpower. Our effort, however, is to extract the knowledge in an automatic way with a lower requirement. Firstly we acquire the knowledge with obvious features manually and build the knowledge base to be the training and testing corpus for automatic extraction. In this article we introduce our guideline for biographic knowledge acquisition from Electronic Encyclopedia of China, including the acceptance and rejection of knowledge items, the formalization of simple knowledge and simple knowledge group and the acquisition of dominant and complicated knowledge.

Keywords: Biographic Knowledge Acquisition, specification

[1] 本文工作得到国家自然科学基金（60272055）、国家 863 计划（2001AA114111）和教育部科学技术研究重点项目（00128）的资助。

0 引言

自然语言文本知识提取是自然语言处理技术的一个的重要应用领域，它可以满足未来计算机发展的智能化与知识化等方面的需求。

面向百科词典的知识提取工作现在越来越受到各方面的重视。中科院计算所曹存根博士带领的课题组在中文百科全书知识提取方面已经作了大量深入的工作[1]。Richard Hull 和 Fernando Gomez 在英文百科辞典知识提取方面工作的特点是基于大容量的语言知识库，计算机自动进行部分分析、语义解释和推理[2]。这两种方法的核心是人工提取知识，提取过程中加入了人的理解，因此能全面深入地挖掘出文本中的知识，所建立的知识库质量较高。但他们共同的缺陷是需要高水平人力的大量投入，成本高，耗时长， 难以大规模地实用化，同时要做到语言知识库的一致性和规范化也很困难。

知识提取是一项高度智能性的工作。即使处理对象是比较规范的百科辞典，也不能指望计算机全自动地完成，必须人机结合。但是，人机结合的结合点在哪里，哪些工作由人做，哪些工作由机器做，如何充分发挥人和机器各自的长处，避免各自的短处，则是很值得研究的。从这种理解出发，我们进行了另一种方法的尝试，即尽量由计算机自动提取百科全书文本中的知识，人工投入尽量地小。这样便可以大量地处理百科全书文本，高速度地建立大规模的知识库。当然，由于计算机的特长是匹配、统计、计算，而在理解能力方面远不如人，因此只能提取显性的、在统计意义具有规范性的知识。

在我们的工作方案中，人工投入还是需要的。主要是人工模仿机器进行这类知识的提取，以便积累经验、摸索规律，设计自动提取知识的算法；此外，人工提取知识的结果将用作机器自动提取算法的训练集和评测集；最后，人工提取得到的知识库也可以直接为用户提供知识服务。

为了达到以上目的，人工提取知识的工作必须规范化。提取知识点的范围和提取结果的表达形式都应同将来计算机自动提取的结果相一致。本文的目的就是介绍人物传记范围内我们的工作规范。

1 人物传记知识提取规划

1.1 工作流程

（1）人物传记词条的提取

首先需要把人物传记词条从百科全书的全部词条中挑选出来。我们从《中国大百科全书》（下面简称为《全书》）55 卷文本共计 78347 个条目中抽取出人物传记 14633 条。

（2）人物传记词条释文中知识的人工提取（详见下文）

（3）知识的存放：《全书》人物传记知识的数据采用成熟的关系数据库形式详细描述人物及其信息属性之间的二维关系。

目前我们已完成纺织、轻工等 9 卷的人物传记知识的人工提取工作，现在人物传记知识提取的各项工作进展顺利。

1.2 制订加工规范的基本思路和知识提取的基本原则

我们首先解释与知识提取相关的术语和约定。

- 简单知识：主要指可直接用对象文本中的一个词或词组来表示的知识。
- 简单知识组：由确定个数的简单知识项组成的知识。例如，人物传记中，多数“职务”知识由任职起始时间、任职结束时间、任职单位、职务名称这四个简单知识项组成。“学位”、“奖项”、“作品”也往往是简单知识组。
- 复杂知识：主要是指包含信息量较多、无法用简单的词汇来描述，只能整体地用句组或段落表示的知识。人物传记中生平、事迹、主要思想等属于复杂知识。
- 形式特征：常见的形式特征有三种。一是条目释文中提示知识的词语。例如，“1829年7月16日生于戈里齐亚”中，“生于”提示其前面的时间短语为出生时间，后面的地点短语为出生地点。二是词语的语义类，如上例中的时间短语“1829年7月16日”和地点短语“戈里齐亚”。三是位置信息，如国籍一般出现于首句的首部，身份出现在首句尾部。
- 显性知识：可直接从原文得到的知识。如“1829年7月16日生于戈里齐亚”中，“1829年7月16日”为出生时间，“戈里齐亚”为出生地点。
- 隐性知识：需要推理才能得到的知识。例如，由“1824年被选入英国皇家学会，次年任皇家研究院的实验室主任。”可知，任皇家研究院实验室主任的时间是1825年，而且这个“皇家研究院”是英国的。但这需要推断“次年”与前面“1824年”的关系、“皇家研究院”与前面“英国皇家学会”的关系才能得知。

我们的目标是计算机自动提取知识。从目前自然语言处理的能力考虑，我们确定提取的对象是具有明确形式特征的、显性的简单知识和复杂知识。也即，暂不提取没有形式特征的知识，以及需要推理才能得到的隐性知识。

人物传记文本中需要提取的知识项目有：姓名、外文名、身份、国籍、民族、原籍国、时代、出生时间、出生地点、死亡时间、死亡地点、籍贯、生平、事迹、学位、任职、奖项、作品、主要思想、又名、字、号、其他名。其中简单知识有姓名、外文名、身份、又名、字、号、其他名、国籍、民族、原籍国、时代、出生时间、出生地点、死亡时间、死亡地点、籍贯。简单知识组项目包括：学位、任职、奖项、作品。复杂知识包括：生平、事迹、主要思想。另有所在条目的两项元数据：卷名和文本号。

2 《全书》人物传记知识库知识提取加工规范

本节给出我们目前从事《全书》人物传记知识提取的基本加工规范。其中简单知识组的构成比较复杂，分为理论规范和操作规范两部分。理论规范的形式比较简单，是知识库最终完成时的形式，便于检索和推理。操作规范的形式比较复杂多样，同原文中自然语言的表述形式比较接近，便于手工采集或机器采集。制订操作规范的目的主要是为了提高知识提取工作的工作效率，同时也是为了保证工作的正确性。符合操作规范的知识表示形式可以用软件自动地转换成符合理论规范的形式。限于篇幅，本文只介绍理论规范。

2.1 简单知识

⑴**姓名**：包括汉族人名和其他民族人名的汉字名或汉译名。一般就是词条本身，不包括别名、字、号等。如“陈第”、“阿斯科里，G.I.”（语言文字卷）

⑵**外文名**：用外文拼写的姓名。一般在词条的后面，紧邻词条。如“奥尔波特，F. Floyd Allport”（奥尔波特，F. 社会学卷）提取为：“Floyd Allport”。

⑶**身份**：标志《全书》人物的总括性概括。介绍人物的国籍、原籍国、职业等信息，往往

是文本的首句。通常有“**家”等提示性词语。如“阿斯科里，G.I. 意大利语言学家。”（阿斯科里，G.I. 语言文字卷）提取为：“意大利语言学家”。

⑷**国籍、原籍国**：《全书》人物传记中首句首词一般会提到国名，这往往是该人物的国籍。如“阿尔贝蒂，L.B. 意大利建筑师、艺术理论家。”（阿尔贝蒂，L.B. 美术卷）则“国籍”和“原籍国”都应该填“意大利”。但有时候是分开的，例如“美籍华人”，则国籍填“美国”，原籍国填“中国”。

⑸**时代**：文本中表示人物生活或活动年代的时间短语，包括“**世纪”“古代”“现代”“当代”等。如“十九*世纪*伟大的科学家”提取为“十九世纪”，中国人物的传记还会用到朝代等，如“边景昭 中国*明代*画家。”（边景昭 美术卷）提取为“明代”等等。

⑹**出生时间、出生地点、死亡时间、死亡地点**：人物传记中常有人物出生时间和死亡时间的时间短语，表示为年月日序列的形式（前面可能有表示纪年方式的词语如“公元前”等），还有人物出生地点和死亡地点的地点短语，并有“生于”、“在……出生”、“卒于”等词语提示。（示例略）

如果文中没有指明出生时间或死亡时间，则可以使用人物传记文本的开头、在外文名的后面用括号列出的生卒年月。（示例略）

⑺**籍贯**：主要针对中国人名，显性标记如“祖籍”“祖上”“***人氏”“***人”等。如：“陈第 连江（今福建连江）人”（陈第 语言文字卷）提取为：“连江（今福建连江）”。

⑻**又名**：人物的正式名以外的名字。显性标记如“又名”“亦称”等。如“……*又名象乾、赤民。*”（杨□ 社会学卷）提取为“象乾、赤民”。

⑼**字，号**：中国历史人物或当代某些人物特有的字和号。《全书》中提及人物的字、号时一般都会用“字**”“号**”标明。如，“章太炎 *字枚叔，号太炎。*”（章太炎 语言文字卷），则字提取为“枚叔”，号提取为“太炎”。

2.2 简单知识组

在这里我们把简单知识组用巴克斯公式描写出来，目的是为了说明问题更方便。其中，方括号“[ ]”、竖线“|”是公式中的元符号，方括号“[ ]”表示可选项，其内容可有可无，竖线“|”表示在列举的几项中选择一项而且必须选择一项；斜体文字表示非终结符，正体文字和左右圆括号“(”“)”以及顿号为终结符。简单知识组中某一项元素或某几项元素若无法提取出确切的知识信息，可填入星号“*”。

⑽**任职**：：=（*任职起始时间，任职结束时间，任职单位，职务名*）

任职起始时间：：=*时间短语*

任职结束时间：：=*时间短语*

任职单位：：=*单位名*

时间短语：：=[*纪年方式*]*整数* 年 [*整数* 月[*整数* 日]][初|中|末|底|上旬|中旬|下旬]

纪年方式：：=[公元|公元前|*帝王年号*]

《全书》文本中一般有“曾任”“担任”“任……”“职务”“主要职务”等特征词语标示人物的任职情况。

示例：“1941～1943 年*担任*美国物价管理局副局长”（加尔布雷思，J.K 社会学卷）.提取为（1941 年，1943 年，美国物价管理局，副局长）。

⑾**学位**：：=（*获学位时间，授学位学校，学位名*）

获学位时间：：=*时间短语*

授学位学校：：=*学校名*

学位名：：=学士|硕士|博士|荣誉博士|名誉博士|副博士

文本中一般用“获……学位”“学位”“主要学位”等特征词语标示人物的学位情况。

示例:“1953 年在哥伦比亚大学*获博士学位*。”(古尔德纳,A.W. 社会学卷)提取为(1953 年,哥伦比亚大学,博士)。

⑿*奖项*:: =(*获奖时间,奖项名称*)

获奖时间:: =*时间短语*

《全书》文本中往往用“获……奖”“荣获”“获得”“奖项”“获奖”“授予”等词语提示人物的获奖情况。

示例:“1964 年*获美国心理学会杰出科学贡献奖*。”(奥尔波特,G.W. 社会学卷)提取为(1964 年,美国心理学会杰出科学贡献奖)。

⒀*作品*:: =(*合作或独立,工作方式,作品类型,作品名*[(*发表时间*)])

合作或独立:: =合作|独立

工作方式:: =著述|主编|编写|改编|表演|设计|作曲……

作品类型:: =著作|论文|文集|建筑|小说|相声|河北梆子……

发表时间:: =*时间短语*

《全书》文本中一般用“著作”“代表作”“著有”“作品”“表演”“编排”“导演”“写有”等形式特征词语来标示人物传记的作品情况。

示例:“1934 年与人*合著*《帕雷托理论介绍》一书”(霍曼斯,G.C. 社会学卷)提取为(合作,著述,著作,《帕雷托理论介绍》)。

⒁*其他名*:: =(*类别,名称*)

《全书》文本中一般用“笔名”“初名”“原名”“别名”等标明类型。

示例:“……*原名*焕鼎”(梁漱溟 社会学卷)提取为(原名,焕鼎)。

2.3 复杂知识

⑴**生平**:主要指一个人从出生到死亡的全过程。主要包括生卒时间、生卒地点、生活和求学经历、职业经历等。《全书》中典型的生平描述往往具有固定的模式。即时间短语+行为事迹,而且其中的行为事迹不展开叙述,只做最一般的概括,往往在一句话内表述完。常有“生平”“一生”等提示词语。

示例(略)

⑵**事迹**:人物传记条目的释文中除了生平之外的内容都归在“事迹”项目中。这里“事迹”是中性词,既可以指正面的行为活动也可以指反面的行为活动。主要包括:社会活动和学术活动、成就、所获得的荣誉、思想、主张或观点、主要贡献等。《全书》中经常出现的形式特征标记有“事迹”“生平事迹”“著作”“主要著作”“代表作”“主张”“担任”等。

示例:“生平事迹 美国发明家……对机械工业的发展曾产生重大影响。”(惠特尼,E. 纺织卷)该部分即可以提取为“事迹”。

⑶**主要思想**:文本中主要有“思想”“主张”“认为”“指出”“提出”“提倡”等显性特征标记。

示例:“阿尔托主要的创作思想是探索民族化和人情化的现代建筑道路。他认为……适应不同地形、不同朝向、不同景色等等。”(阿尔托,A. 建筑园林卷)该部分即可以提取为“主要思想”。

3 进一步的工作

目前我们的人物传记知识提取工作正在有条不紊地进行。进一步的工作目标是:

- 进一步完善《全书》人物传记数据库的构建流程，完善规范，开发辅助知识提取工具；
- 形成较为统一的，可以满足不同层次知识提取工作的样板；
- 在完善手工知识提取成果的基础上，研制机器自动提取系统。

总之，合理、有效地开展百科辞典知识基础资源建设，推动相关的加工技术和应用技术不断向前发展，是我们的努力的方向。

参 考 文 献

[1] Gu, Fang. and Cao, Cungen.: Biological Knowledge Acquisition From the Electronic Encyclopedia of China, Proc. of ICYCS'2001, 2001, pp.1199-1203.

[2] Richard Hull, Fernando Gomez: Automatic acquisition of biographic knowledge from encyclopedic texts, ExpertSystems with Applications 16(1999), pp.261-270

[3] 许勇，宋柔：基于百科辞典的知识获取系统的研究于实现，首届学生计算语言学研讨会论文集，北京，2002.8.20—8.23。

[4] 俞士汶，段惠明，朱学锋，孙斌：北京大学现代汉语语料库基本加工规范，中文信息学报，2002，16（5）。

[5] 柏晓静，常宝宝，詹卫东，吴拥华：构建大规模的汉英双语平行语料库，“2002 全国机器翻译论坛”会议论文