基于PADL的古代人物简历知识获取*)

郝天永　曹存根
(中国科学院计算技术研究所智能信息处理重点实验室　北京100080)

摘　要　领域文本知识获取是目前人工智能中的一个关键问题。本文探讨如何从人物简介中获取人物知识。由于自然语言技术目前尚不足支持自动的知识获取，某种形式的人机交互或半自动方法是一种可行的折衷方案。本文在总结人物知识描述的特点基础上，提出了一种中间标记语言，它是自然语言到目标知识表示语言的过渡桥梁。同时，我们还介绍使用该方法在宗教古代人物知识获取中的应用。

关键词　知识获取，中间语义标记语言，PADL

Profile Acquisition from Text Based on an Intermediate Semantic Markup Language

HAO Tian-Yong　CAO Cun-Gen
(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

Abstract　Domain knowledge acquisition is one of key problems in artifical intelligence. This paper discusses how to acquiring knowledge from brief introduction of ancient people. Natural language techniques can't support auto-acquisition of domain knowledge perfectly, human-machine interaction or semi-auto-acquisition becomes a moderate method. Based on the summary of features of human knowledge description, this paper puts forward an intermediate semantic markup language, which sets up a bridge between natural language and expression language of object knowledge. We also introduce how to use this method to acquire religious knowledge of ancient people.

Keywords　Knowledge acquisition, Intermediate semantic markup language, PADL

1　引言

最近几年，知识的大规模获取、形式化表示、定性与定量分析、共享和应用已越来越受到人们的重视。主要原因是人们已意识到知识在自然语言理解、语音识别、机器学习、智能教学等研究中具有不可替代的作用。领域专业知识作为知识体系的重要组成部分，受到了人们的关注。

文本知识获取是人工智能中非常重要的研究领域之一[1~4,6~13,15~18,20,22,23~30]。在文本知识获取过程中，有两个不同的方法[2]，第一个方法依赖一般目的的算法来理解自然语言文本并从文本中获取它们之间的联系[11~13,15,25,26]。据报道，这种方法可以从文本中获取相当量的知识，但是，由于自然语言的复杂性，文本知识获取不能够达到彻底的自动化。从某种意义上来说，第二种方法是一种折中的方案，即半自动的方法。知识工程师采用语义标记文本[1,2,7,18,22,23,29]进行文本知识获取。换句话说，整个过程分为人工处理和计算处理两部分。

知识获取的目的是让计算机能"理解"这些知识，我们采用谓词逻辑形式的目标语言，但在实际的获取过程中，知识工程师需要中间语义标记语言方便知识的形式化，中间标记语言架起了自然语言到目标语言的桥梁。尽管现有的中间标记语言很多，但人类知识的学科范围很广，领域知识体系庞大且十分复杂，单从宗教来看，就有人物、神灵、王朝、事件、物品等两百多个子类，这使得在特定的领域中，现有的方法不能结合具体的情况，而显得力不从心。

经过了长期的领域知识形式化，我们积累了大量的形式化经验，结合现有的一些标记语言，我们提出了一种新的中间标记语言—PADL(Procedural and Declarative Language)，并制定了相应的处理规则。PADL已经开始在宗教等诸多领域中进行实践应用，随着处理规则的不断成熟和完善，它将在特定领域的知识获取中发挥越来越重要的作用。

本文第2节介绍古代人物简历知识获取的困难和中间语言标记语言使用的必要性。第3节介绍PADL在古代人物知识获取中的优势。第4节介绍PADL相关处理规则的6个不同部分及其使用方法。第5节介绍从中间标记语言到目标语言的转化。最后总结目前的工作，以及PADL的下一步工作。

2　古代人物简历知识获取的困难

在知识获取的过程中我们发现，特定领域的知识本身有着自身显著的特点，就古代宗教人物而言，就有如下特点：

1)描述方式的多样化：不同的宗教有着自身宗教发展的特色，因而形成了很大的差异性，如中国藏传佛教和中国道教，其描述方式就有显著差异。

2)内容组织的复杂性：宗教悠久的发展历史和人类社会发生了显著的关系，涉及的内容范围很广且复杂，与其它学科(如：天文、地理、军事、经济、化学等)都有密切关系。

3)上下文的相关性：由于人的活动具有过程性、交互性，人物的简历知识同样具有很强内容的上下文相关性，如思想变化、求学经历、任职状况等。

4)内容描述的不完整性：由于历史的久远、记录的不详尽、人类活动的复杂性，很多古代人物并没有翔实完整的记录，而以跳跃性记录为主，这种简历记录的模糊性、不完整性

*)本文工作得到自然科学基金(#60073017和#60273019)和科技部重大基础研究专项(#2001CCA03000和#2002DEA30036)的资助。郝天永　硕士研究生；曹存根　博士生导师。

甚至缺省对于知识获取都造成了很大困难。

5)缺乏特殊词语词典:要做到真正的自然语言理解,必须对自然语言文本能进行无歧义无错误的语义切分,因而就必须建立足够的词典,例如古人名(尤其是外国人名)、地名、事件名、官职名等。而使用中间标记可以获取这些词语。

这些领域知识的描述特点给知识获取造成了很多困难,但这些知识的正确有效获取,对于建立特殊词语词典、过程性知识获取、其它领域的知识获取等都有积极的意义。

就目前来看,一方面,由于自然语言的复杂性,自然语言技术无法做到自动获取;另一方面,类自然语言方法 BKDL 也有自身局限性[32],BKDL 可以很好地描述分类知识,但却难以描述过程性的人物简历知识。

基于简历知识获取的诸多困难,而自然语言技术和 BKDL 又不能很好地解决这一问题,我们采用了折中的方法,使用中间标记语言。知识工程师通过中间标记语言对文本知识进行处理,然后将中间标记语言转化成目标语言。中间标记语言使知识工程师的知识获取更具有灵活性,可以获取特殊词汇来建立词典,又可以很好地处理过程性知识,还具有易于使用、易于处理多谓词等其它优点。

3 PADL 在古代人物知识获取中的优势

在大量领域形式化的基础上,我们提出了"PADL"的中间语义标记语言。这种方法允许知识工程师直接在原文本的基础上进行手工形式化处理,在减少手工处理的工作量的同时尽量保证形式化文本的语义正确性、完整性、有效性,我们认为该方法有以下显著优点:

1)易于使用,提高工作效率。知识工程师使用"PADL"对自然语言形式化时,直接在原文进行文字的组合处理,而不需要转化成谓词逻辑的形式,因而就大大减少了手工工作量;处理时,PADL 结合原文本进行组合,而不需要大量的修改原文本,提高了工作效率。如:原文本:"1596年定居恺撒利亚",处理后:"#1596年 !c 定居 恺撒利亚"。

2)结构清晰。"PADL"使用关键字组合的方式,连接符采用"+"等易于理解的符合,并使用分隔符使得整体结构清楚明了。如:" # 622年 !c 随+出走 (穆罕默德;麦地那)",很容易理解,"622年随穆罕默德出走至麦地那"。

3)有助于知识工程师把握知识间的联系。很多知识之间是有上下文、有联系的,形式化将知识分隔开,极易丢失信息,"PADL"充分考虑到了知识的上下文关系,使用特定语义分隔符保留这些信息,并且"组合"的方法,使知识工程师更易于浏览、处理、整理这些关系。如:"#1663年 !c 反对 改革 &(原因-结果)1664年 !c 被流放到 梅津"。

4)易于处理"多谓词"情况。自然语言极其复杂,经常会出现多个谓词,"PADL"采用组合方式容易处理不同情况的"多谓词",并且处理的结果以组合的方式处理,结构十分清晰。如:原文本:"631年代表穆罕默德率信徒赴麦加朝觐",处理后:"#631年 !c 代表+施事 (穆罕默德;率信徒赴麦加朝觐)"。

5)较好处理知识缺省和不完整。很多古代宗教人物的描述并不完整,甚至缺失,例如时间、对象等的缺失,"PADL"对时间、主体等进行了详细的分类并给出了相应的处理方法,能较好地进行正确处理。如:事件时间的缺失。通过使用符合"#"(含有时序关系),当从"PADL"转化成目标语言的时候,就会处理成 Time_after(e1,e2)等,从而找出事件发生的时间段。

6)方便向目标语言的转化。由于"PADL"采用标记组合的方式,而不是固定关系的方法,因而在向目标语言转化的时候,就具有很大的灵活性,直接可以对组合的标记进行转化形成谓词逻辑。如:PADL:"#631年 !c 代表+施事 (穆罕默德;率信徒赴麦加朝觐)",目标语言:"代表+施事(631年;!c;穆罕默德;率信徒赴麦加朝觐)"。

4 PADL 及其相关处理规则

PADL 方法由符号、时间、主体、谓体、宾体和群六个部分组成,如图1所示。它采用标记组合的方式灵活处理古代人物知识。从语言结构来看,主要分为四大部分:时间、主体、谓体和宾体,四部分紧紧结合在一起形成对知识的完整描述。这里主体、谓体和宾体不同于主语、谓语和宾语,它们是"PADL"中的一个结构组成部分,用来区分整体结构的各个部分,由于其结构酷似主谓宾的语法结构,因此命名为主体、谓体和宾体。表1结合实例给出了槽方法和"PADL"方法的对比,并给出了"PADL"的四大结构组成部分。

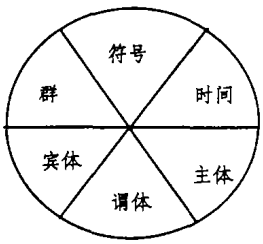

图1 PADL 的组成部分

表1 PADL 的结构

原文	阿依沙	9岁时与穆罕默德在麦地那结婚
PADL	#阿依沙9岁时 !c 与结婚+地点 (穆罕默德;麦地那)	
	时间:	阿依沙9岁时
	主体:	!c
	谓体:	与结婚+地点
	宾体:	(穆罕默德;麦地那)

4.1 符号部分

符号部分是"PADL"的第一部分,也是最基础的部分,它定义了该中间标记语言中所需要的所有符号,见表2,这些符号有着不同的作用,可以使手工形式化更具可操作性,其目的是方便计算机的识别和处理。

表2 PADL 的符号表

符号	符号解释
#	有时序的知识分隔符,用以区分不同时间段的事件
&	有时序的知识连接符,用以连接同一时间段的不同事件
&()	有时序和关系的知识连接符,& 表示时序,() 中含有知识之间的关系
Backspace	主体、谓体、宾体的划分符,用以区分和识别主体、谓体和宾体
!c	主体标识符,指主体是当前框架描述的对象
!o	主体标识符,指主体不是当前框架描述的对象
+	谓体内部组合符,用以对谓体单元进行组合
()	宾体标识符,其内部是宾体单元;群内部事件、活动、状态等的表示符
;	宾体内容体的区分符,用以进行区分不同的宾体单元

4.2 时间部分

时间部分是"PADL"的第二部分,由于古代人物的简历知识有很多的缺失、不完整等,其中一个重要的方面就是时间

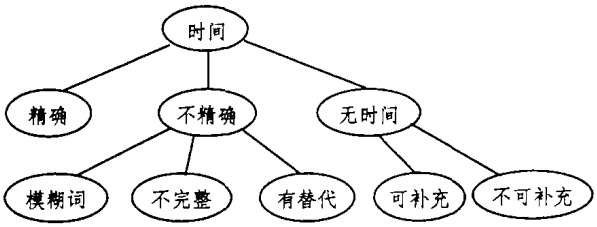

图2 领域形式化时间的分类结构

缺失和不完整，大量的原始时间中含有不精确现象，例如："二十三年，死于西藏"(二十三年是指什么历法，什么年号?对应公历那一年?)；"18岁学律法"(谁18岁?)等等问题。基于大量的形式化工作，我们分类总结了类似的问题，为方便手工形式化工作，将时间按照精确程度进行分类，如图2。

根据按时间精确程度的分类结果，我们制定了"PADL"相应的处理规则，并结合实例解释说明，见表3。

表3 PADL 的时间处理规则表

时间类型			处理方法		实例	实例的处理
有时间	精确型		直接保留原时间		1609年任伦敦主教	#1609年 !c 担任 伦敦主教
有时间	不精确型	不精确词语	使其尽可能的精确	结合上下文时间综合描述	…后创建博多寺	#1058年后 !c 创建 博多寺
有时间	不精确型	意义不完整	使其尽可能的精确	意义补充完整	…9岁在麦地那与穆罕默德结婚	#阿依沙9岁时 !c 与结婚+地点(穆罕默德;麦地那)
有时间	不精确型	替代型时间	使其尽可能的精确	结合上下文，用精确时间替换	当年被选为米兰主教	#374年 !c 被选为米兰主教
无时间		可补充型	结合上下文内容补充相应的时间		定居康科德	#爱默生回国后 !c 定居康科德
无时间		不可补充型	上下文找不到时间，则空缺		主持苏拉律法学院	#!c 主持 苏拉律法学院

4.3 主体部分

主体部分是"PADL"语言的第三部分，对主体进行分类总结，并用相应的规则符号进行替代，可以方便计算机进行识别和处理，更可提高工作效率。经过大量的总结，我们发现，以自然语言方式组织的信息主体形式多种多样，替代、隐含甚至缺失主体情况非常多，例如："十三年夏，辗转至河州。一说来自西域，可能是阿帕克和卓的同路人、在传教活动中学习汉语，攻读医儒释道之书。"这样的句子有时难以理解其主体部分，根据句子的结构特征进行分类总结，如图3所示。

形式化中，针对不同的主体特点按照符合表使用规则符号进行标识，这些符合可以方便计算机进行识别和处理，并为自动获取方法积累经验。表4是按照主体分类的不同制定的主体领域形式化处理规则表。

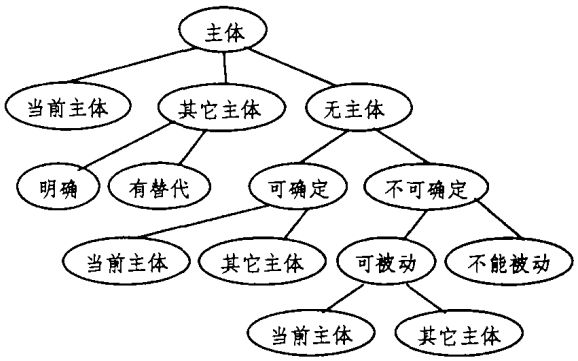

图3 领域形式化主体的分类结构

表4 PADL 的主体处理规则表

主体类型			处理方法	符号	实例	实例的处理
当前主体(1)	主体是当前框架的描述对象		直接用相应符号声明	!c	1883年应缅生于哥伦比亚	#1883年 !c 出生地点 哥伦比亚
其他主体(2)	主体不是当前框架描述对象	明确给出了主体	相应符号+主题内容	!o	葡萄牙当局加以阻挠	# !o 葡萄牙当局 阻挠 卜弥格
其他主体(2)	主体不是当前框架描述对象	给出了主体的替代词	替换成精确独立的主体	!o	其父亡于893年	# !o 阿米尔之父 死亡
无主体	主体缺失	根据上下文能确定合适主体类型	主体是当前主体，则填充并转(1)	!c	1823年消灭鲁萨	# 1823年 !c 消灭鲁萨
无主体	主体缺失	根据上下文能确定合适主体类型	主体是其它主体，则填充并转(2)	!o	对穆阿维叶施加压力	# !o 阿布杜·巴哈 对施压 穆阿维叶
无主体	主体缺失	根据上下文不能确定合适主体类型	变被动形式，主体是当前主体，则填充并转(1)	!c	…589年杀死卡鲁森	# 589年 !c 被杀
无主体	主体缺失	根据上下文不能确定合适主体类型	变被动形式，主体是其它主体，则填充并转(2)	!o	曾盗用张理著作	# !o 张理的著作曾被盗用
无主体	主体缺失	根据上下文不能确定合适主体类型	不能变被动，则说明知识本身存在错误，故不作处理，直接遗弃			

4.4 谓体部分

谓体部分是PADL语言的第四部分，其体现了组合的思想，即：关键词和侧面通过标记进行组合。关键词(即中心词)只有一个，这样就可以用一个标记和多个侧面进行组合(组合是有序的)，形成谓体。谓体单元是谓体的基本组成部分，它有标记和侧面两种，其间用"+"连接。谓体单元组合方法因实际信息的不同而有所不同，有"标记"、"标记+侧面"、"侧面+标记"等几种，不同的组合方式有相应的处理方法，见表5。

表5 PADL的谓体部分表

谓体类型			处理方法	实例	实例的处理
一般情况	关键词		提取标记	1625年为查理一世加冕	#1625年 !c 为加冕 查理一世
	关键词+侧面1+…侧面 n(n>=1)	关键词+侧面	提取标记(标记提取方法见表6)和侧面,按左边类型进行组合,中间用"+"隔开	1597年获巴利奥学院神学博士	#1597年 !c 获学位+授予单位(神学博士;巴利奥学院)
		关键词+侧面1+…侧面 n(n>1)		1597年在法国被教皇任命为博格堪普大主教	#1597年 !c 被任命为+任命人+地点(博格堪普大主教;教皇;法国)
特殊情况	侧面1…侧面 m+关键词(m>=1)	侧面+关键词	提取标记(标记提取方法见表6)和侧面,按左边类型进行组合,中间用"+"隔开	622年和穆罕默德一起出走	#622年 和一起+出走(穆罕默德;)
		侧面1+…侧面 m+关键词(m>1)		839年与何采一起编写《冷语》	#839年 与+编写(何采;《冷语》)
	侧面1…侧面 m+关键词+侧面1+…侧面 n(m>=1)(n>=1)			622年和穆罕默德一起出走麦地那	#622年 跟随+出走+出走地点(穆罕默德;麦地那)

其中"关键字"可以用来描述对象的属性,或与其他对象的关系,从这个角度而言,可以分为关系标记和属性标记。在实际的文本中,标记的出现并不都是精确的,见表6。根据这些情况,我们制定了相应的处理规则。

表6 谓体关键字的类型及处理方法表

关键词类型		说明和处理方法	实例	实例的处理
关系	不完整	提取关键词语,将其补充完整	1609年任利奇菲尔德主教	#1609年 !c 担任 利奇菲尔德主教
	词义模糊	提取关键词语,根据语境精确化	正一则总辅六部	# !o 正一 补充解释 六部
	缺失	根据语境总结出"标记",找出所有情况,方便自动获取	1597年牛津大学神学博士	#1597年 !c 获学位+授予单位(神学博士;牛津大学)
属性	不完整	提取关键词语,将其补充完整	阿布·伯克尔早期经商	#!o 阿布·伯克尔 早期职业 经商
	词义模糊	提取关键词语,根据语境精确化	阿布·伯克尔是四大正统哈里发之一	#!o 阿布·伯克尔 是成员 四大正统哈里发
	缺失	根据语境总结出"标记",找出所有情况,方便自动获取	李四八尺有余	#!o 李四 身高 大于八尺

谓体建立的关键是如何寻找标记和侧面,遵循的原则是,先寻找动词,优先考虑将动词作为标记,如果无动词则考虑形容词和副词。但在实际的信息中,可能出现大量的多动词,见表7,我们根据动作的类型将动词作为关键字进行处理,其它动词作为侧面。

表7 多动词谓体处理规则表

类型	说明与处理	实例	实例的处理
任务类	带有某种任务相关的动作,将其作为标记,其余动词为侧面	1833年,被派往也门作战	#1833年 !c 被派往+任务(也门;作战)
目的类	带有目的性的动作,将其作为标记,其余动词为侧面	1677年,转移到巴黎工作	#1677年 !c 转移到+目的(巴黎;工作)
施事类	当前主体是动作的发起者,将首个动词作为标记,其余为侧面	631年代表穆罕默德率信徒赴麦加朝觐	#631年 !c 代表+施事(穆罕默德;率信徒赴麦加朝觐)
受事类	当前主体是动作的承受者,将首个动词作为标记,其余为侧面	474年,被关入监狱折磨	#474年 !c 被关入+受事(监狱;折磨)
经事类	当前主体不是事件的发起者和承受者,而是经历者。将首个动词作为标记,其余为侧面	858年 赴英国经历了英国宗教的一系列变动	858年 !c 赴地点+经事(英国;英国宗教的一系列变动)

4.5 宾体部分

宾体部分是PADL语言的第五部分,也是其形式结构的最后一部分。宾体单元是宾体的基本单位,不同的宾体单元与谓体单元一一对应起来,组成宾体。如:"#1597年 !c 获学位+授予单位(神学博士;牛津大学巴利奥学院)",其中宾体是:(神学博士;牛津大学巴利奥学院),而"神学博士"和"牛津大学巴利奥学院"则是两个宾体单元,并与谓体单元"获学位"和"授予单位"一一对应起来。表8给出了"PADL"宾体部分的处理规则。

宾体的处理还遵循以下规则:当要处理的信息无宾体,并且根据上下文找不到合适的宾体时,允许宾体空缺;如果宾体中只有一个宾体单元,则不必用"()"标识符;当宾体中的宾体单元超过一个,则用";"隔开;宾体单元允许空缺,但原则上空缺数目不能大于一个,且小于宾体单元总数。

宾体是与谓体一一对应的,每一个宾体单元对应于一个谓体单元(标记或侧面),这样的对应更有助于计算机的识别

和处理。这种一一对应并不是严格的，宾体允许空缺，但是一旦拥有宾体单元，就必须严格满足一一对应关系。

表8 PADL的宾体部分表

宾体类型		处理方法	实例	实例的处理
有宾体	单个宾体单元	将单个宾体单元作为宾体	#早年学过哲学、罕百里教法和圣训	#阿布早年 !c 学过 哲学，和 罕百里教法，和 圣训
	多个宾体单元	将其用宾体表示方法表示(;)	#1597年获牛津大学巴利奥学院神学博士	#1597年 !c 获学位+授予单位（神学博士;牛津大学巴利奥学院）
无宾体	上下文分析，可以补充的	根据上下文，补充合适的宾体	（他很想学律法），于是21岁开始学习	#王担21岁时 !c 开始学习 律法
	不能补充的	空缺宾体	18岁才开始学习	#18岁 !c 开始学习

定理1 |谓体组合单元|=|宾体单元|(有宾体)。

定理2 “谓体组合单元”到“宾体单元”是一一映射关系(有宾体)。

4.6 并列群、时序群和关系群

在形式化的过程中我们发现，很多知识是有关系的，是相互关联着的。在某些特定领域，这种关联性尤为明显，例如宗教人物、宗教事件活动等，绝大多数以并列群、时序群和关联群的形式出现。

我们把在同一时间发生的多个事件、动作、属性或状态等的集合叫作一个并列群。群内部事件、动作和属性等发生在同一时间或时间段，其有显著的同时或并列关系。例如：“972年在蓬贝迪塔律法学院执教，并协助其父”，在972年这个时间段同时存在两个活动，且有并存关系，这两个并列的活动构成了一个并列群。我们总结了并列群的情况，并给出了相应的处理方法，见表9。

我们把仅有时间顺序的多个事件、动作、属性或状态等的集合叫作一个时序群。群内部之间的关系仅仅是时序关系，并无其他语义上的内在联系。例如：“1596年定居恺撒利亚，任拉比学院院长”，在1596年这个时间段中，发生了“定居恺撒利亚”和“任拉比学院院长”两个事件，而它们之间仅有时序关系，这两个事件构成一个时序群。我们总结了时序群的情况，并给出了相应的处理方法，见表10。

我们把其中有语义联系的那些事件、动作、属性或状态等的集合叫作一个时序关系群。群内部有语义上的联系，而且常常有时序关系。例如：“十月革命胜利后，曾参加反对苏维埃政权的活动，1922年5月被捕”，“参加反对苏维埃政权的活动”和“被捕”两个事件之间是有内在联系的，这两个事件就构成了一个关系群，而且这两个事件之间有时序关系。我们总结了关系群的情况，并给出了相应的处理方法，见表11。

由于形式化的知识都以自然语言的方式存在，其中的内容含有大量语义上的联系，因而研究“语境”是十分重要的。手工形式化时，是人工地进行语义理解，语义和语境的理解多取决于知识工程师对专业知识的理解和掌握。在对时序群和关系群进行处理时，我们遵循“保留关系”的原则，不对关系群进行拆分，保留之间的联系，用不同的符号进行说明。

表9 “并列群”处理方法表

类型：同一时间段的并列群		处理方法：使用“+(标记)+”连接不同事件、活动或状态等，标记表示并列关系，根据原文，可以是“同时”、“并”、“又”等。
举例	原文	从私人导师接受传统宗教教育，同时在纽约市学院获得学位。
	处理	# !c 从+接受教育（同时）获得学位于（私人导师;传统宗教教育;纽约市学院）

表10 “时序群”处理方法表

类型：同一时间段内的时序群		处理方法：使用“&”分隔符连接时序事件、状态等
举例	原文	1903年来中国上海徐家汇，攻读哲学，辗转杭州。
	处理	# 1903年 !c 来到 & 攻读 & 辗转（中国上海徐家汇;哲学;杭州）
类型：不同时间段内的时序群		处理方法：使用“#”分隔符作为不同的知识条分隔开来
举例	原文	1914年起任教于杜宾根大学，1936年曾在美国讲学
	处理	#1914年起 !c 任教于 杜宾根大学 #1936年 !c 曾讲学地点 美国

表11 “时序关系群”处理方法表

类型：同一时间段内的关系群		处理方法：使用“&(标记)”，标记根据原文，可以是表转折的“但”、“却”，也可以是表因果的“所以”、“故”、“因此”等，表示不同关系。
举例	原文	1923年5月被主教公会决定撤销其牧首职务，但本人不接受
	处理	#1923年5月 !c 被决定+撤销职务 &(但)是否接受 态度(主教公会;牧首;否)
类型：不同时间段的关系群		处理方法：使用“&()”分隔符表示群内的事件、状态等有上下文联系，且之间存在时序关系，()里面表示之间的关系类型，如：“状态－导致事件”
举例	原文	十月革命胜利后，曾参加反对苏维埃政权的活动，1922年5月被捕。
	处理	# 十月革命胜利后 !c 曾参加活动 反对苏维埃政权 &(事件－结果)1922年5月 !c 被捕

5 从中间标记语言到目标语言的转化

5.1 基于规则的算法

知识获取的任务是为信息系统或者专家系统获取知识，建立起健全、完善、有效的知识库，以满足求解领域的问题的需要。其目的是让计算机能"理解"和"掌握"这些知识，因而必须要求有一种最终语言，而这种语言采用谓词逻辑的形式，这就是目标语言。

知识工程师有着自己的思维习惯，并十分符合自然语言的组织形式，如果知识工程师把自然语言直接转化成目标语言，就会大大增加其工作量，并且难以控制形式化的质量，借助中间标记语言，则可以有效地提高手工形式化的效率并且方便编译成最终需要的目标语言。

由于目标语言是谓词逻辑的形式，因此中间标记语言必须保证能够合理地转化成逻辑的形式。"组合符合"的转化形式很简单，首先判断知识条中是否含有"#"，有则表明是有时序关系的过程知识，转化成逻辑 after(T2,T1)，before(T1,T2)的形式，两者是逻辑等价的。同理，"&"也是含有时序关系的，但"&"也可以含有语义关系，如："因一果"这就需要在目标语言中进行定义，从而转化成 after(T2,T1)and Cause-Effect(E1,E2)的形式。

表12 部分从"PADL"中间语言到目标语言的转化

"PADL"中间语言	目标语言
#T !c A B	A(T,O,B)
#T !c A+B(C; D)	A+B(T,O,C,D)
#T1 !c A+B(C; D) #T2 !c E F	A+B(T1,O,C,D)and E(T2,O,F)and after(T2,T1)
#T !c A+B & E (C; D; F)	A+B(T,O,C,D)and E(T,O,F)and after (Time(E),Time(A+B))
#T1 !c A B &(Cause-Effect)T2 !c C D	A(T1,O,B)and C(T2,O,D)and after (T2,T1)and Cause-Effect(Event(A), Event(C))

5.2 一个具体例子

基于从 PADL 到目标语言的转化算法，我们给出一个完整的古代宗教人物实例，如图4，然后使用 PADL 将其转化成中间语义标记语言，如图5，最后将其转化成目标语言，如图6、图7，两者都是谓词逻辑的形式，区别在于，图7是过程性知识，是有时序关系的。

[人物]阿布·俄拜德（Ab[u1] 'Ubayda b. al-Jarr[a1]h，？—639）伊斯兰教早期将领。生于麦加古来氏部落哈里斯家族。7世纪20年代中叶，迁移到埃塞俄比亚。后参加白德尔和伍侯德战斗。曾被派往也门任教。穆罕默德逝世前回到麦地那。和阿布·伯克尔、国麦尔等人参加推举哈里发的"赛基发会议"。633年率第一路军远征叙利亚；成为阿拉伯军统帅。在叙利亚死于瘟疫。

图4 古代宗教人物实例原文

总结与下一步工作 本文基于领域知识的形式化，结合长期积累的大量形式化经验，提出了一种新的中间语义标记语言－PADL，在此基础上，制定了相应的形式化处理规则，在对古代宗教人物简历的知识获取中，取得了良好的效果。目前，PADL 还有很多不完善的地方，需要进一步的修整和改进，其修订工作将长期进行下去，一方面把处理规则往细节方面发展，使之真正成为"准则"；另一方面将向抽象概括的方面发展，以便能和更多的领域结合起来，使其能在更多更广的领域范围内发挥作用。

```
defframe 阿布·俄拜德
{
!c    阿拉伯名 Ab[u1] 'Ubayda b. al-Jarr[a1]h
!c    出生时间 不详
!c    死亡时间 639年
!c    是一个 伊斯兰教早期将领
!c    出生地点 麦加
!c    出身 古来氏部落哈里斯家族
# 7世纪20年代中叶 !c 迁移到 埃塞俄比亚
# 后 !c 参加战斗 白德尔战斗，和 伍侯德战斗
# !c 曾被派往+任务 （也门；任教）
# 穆罕默德逝世前 !c 回到 麦地那
# !c 和+参加 （阿布·伯克尔，和 国麦尔 等人；
  推举哈里发的"赛基发会议"）
# 633年 !c 率+远征 （第一路军；叙利亚） &
# !c 成为 阿拉伯军统帅
!c    死亡地点 叙利亚
!c    死亡原因 瘟疫
}
```

图5 使用 PADL 转化后的中间标记语言

```
阿拉伯名（阿布·俄拜德，Ab[u1] 'Ubayda b. al-Jarr[a1]h）
出生时间（阿布·俄拜德，不详）
死亡时间（阿布·俄拜德，639年）
是一个（阿布·俄拜德，伊斯兰教早期将领）
出生地点（阿布·俄拜德，麦加）
出身（阿布·俄拜德，古来氏部落哈里斯家族）
死亡地点（阿布·俄拜德，叙利亚）
死亡原因（阿布·俄拜德，瘟疫）
```

图6 转化后的目标语言

```
事件表示：
e1＝时间＋迁移到(7世纪20年代中叶，阿布·俄拜德，埃塞俄比亚)
e2＝参加战斗(阿布·俄拜德，白德尔战斗,和 伍侯德战斗)
e3＝曾被派往+任务(阿布·俄拜德，也门，任教)
e4＝时间＋回到(穆罕默德逝世前，阿布·俄拜德，麦地那)
e5＝和+参加(阿布·俄拜德，阿布·伯克尔,和 国麦尔 等人,
    推举哈里发的"赛基发会议")
e6＝时间＋率+远征(633年，阿布·俄拜德，第一路军，叙利亚)
e7＝成为(阿布·俄拜德，阿拉伯军统帅)
事件时间顺序：
e2:  after(e2, e1) and before(e2, e3)
e3:  after(e3, e2) and before(e3, e4)
e5:  after(e5, e4) and before(e5, e6)
e7:  after(e7, e6)
```

图7 转化后的目标语言

参 考 文 献

1 Bowden P R, Halstead P, Rose T G. Extracting Conceptual Knowledge from Text Using Explicit Relation Markers. In: N. Shadbolt, K. Ohara, and G. Schreiber, eds. Lecture Notes in Artificial Intelligence 1076, Springer-Verlag, Berlin, 1996. 147～162

2 Cao C G. Medical Knowledge Acquisition from Encyclopedic Texts. Lecture Notes In Computer Science 2101, Springer-Verlag, Berlin, 2001. 268～271

3 Chaudhri V K, Farquhar A, et al. The Generic Frame Protocol 2.0; [SRI International Technical Report]. 1997

4 Delisle S, Barker K, Copek T, Szpakowicz S. Interactive Semantic Analysis of Technical Texts. International Journal of Computational Intelligence, 1996, 12: 273～306

5 The Encyclopedia of China, Medical Volume. the Encyclopedia of China Press, 1991

（下转第222页）

tion. Amsterdam: North-Holland, 1979. 3～18
5 Pawlak Z. Granularity of knowledge, indiscernibility and rough sets. Proceedings of 1998 IEEE Intl. Conf. on Fuzzy Systems, 1998. 106～110
6 Polkowski L, Skowron A. Towards adaptive calculus of granules. In: Proc. of 1998 IEEE Intl. Conf. on Fuzzy Systems, 1998. 111～116
7 Skowron A, Stepaniuk J J. Information granules and approximation spaces. Manuscript, 1998
8 Lin T Y. Granular computing on binary relations I: data mining and neighborhood systems, II: Rough set representations and belief functions. In: Polkowski L, Skowron A, editors. Rough sets in knowledge discovery 1. Heidelberg: Physica-Verlag, 1998. 107～140
9 Yao Y Y. Granular computing using neighborhood systems. In: Roy R, Furuhashi T, Chawdhry, PK, eds. Advances in soft computing: engineering design and manufacturing. London: Springer-Verlag, 1999. 539～553
10 Yao Y Y. Rough sets, neighborhood systems, and granular computing. In: Proc. of the 18th Intl. Conf. of the North American Fuzzy Information Processing Society. IEEE Press, 1999. 800～804
11 Klir GJ. Basic issues of computing with granular computing. In: Proc. of 1998 IEEE Intl. Conf. on Fuzzy Systems; 1998. 101～105
12 张文修，吴伟志，梁吉业，李德玉. 粗糙集理论与方法. 北京：科学出版社，2001
13 卜东波，白硕，李国杰. 聚类/分类中的粒度原理. 计算机学报，2002，25(8)：800～806
14 苗夺谦，范世栋. 知识的粒度计算及其应用. 系统工程理论与实践，2002，1：48～56

(上接第174页)

6 Feng Y, et al. A Handbook of TCM Prescriptions. Guizhou Science and Technology Publishing House, 1999
7 Gu F, Cao C G. Biological Knowledge Acquisition from the Electronic Encyclopedia of China. In: Proc. of the 16th Intl. Conf. for Young Computer Scientists, 2001. 1199～1203
8 Gomez F. Acquiring Knowledge about the Habitats of Animals from Encyclopedic Texts. In: Proc. of the Workshop for Knowledge Acquisition, 1995. 1～22
9 Gomez F, Hull R, Segami C. Acquiring Knowledge from Encyclopedic Texts. In: Proc. of the 4th Conf. on Applied Natural Language Processing, 1994. 84～90
10 Hahn U, Romacker M, Schulz S. Discourse Structures in Medical Reports - Watch Out! The Generation of Referentially Coherent and Valid Text Knowledge Bases in the MEDSYNDIKATE System. International Journal of Medical Informatics, 1999, 53: 1～28
11 Hahn U, Schnattinger K. Deep Knowledge Discovery from Natural Language Texts. In: D. Heckerman, H. Mannila, D. Pregibon and R. Uthurusamy, eds. Proc. of the 3rd Intl. Conf. on Knowledge Discovery and Data Mining, Menlo Park, CA: AAAI Press, 1999. 175～178
12 Hahn U, Klenner M, Schnattinger K. Learning from Texts: A Terminological Metareasoning Perspective. In: S. Wermter, E. Riloff and G. Scheler, eds. Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing. Lecture Notes In Artificial Intelligence 1040, Springer-Verlag, Berlin, 1996. 453～468
13 Hahn U, Romacker M. Content Management in the SYNDIKATE System - How Technical Documents Are Automatically Transformed to Text Knowledge Bases. IEEE Transactions on Data & Knowledge Engineering, 2000, 35: 137～159
14 He X D. The Chinese Materia Medica (China Academy Press, 1998)
15 Hull R, Gomez F. Automatic Acquisition of Biographic Knowledge from Encyclopedic Texts. International Journal of Expert Systems with Applications, 1999, 16: 261～270
16 Kazawa K, Fujimoto K, Matsuzawa K. Attribute Dependency Acquisition from Formatted Text. In: Proc. of the 3rd Intl. Conf. on Knowledge-Based Intelligent Information Engineering Systems, 1999. 464～468
17 Lapalut S. Text Clustering to Help Knowledge Acquisition from Documents. In: N. Shadbolt, K. Ohara and G. Schreiber, eds. Advances In Knowledge Acquisition. Lecture Notes in Artificial Intelligence 1076, Springer-Verlag, Berlin, 1996. 115～130
18 Lei Y X, Cao C G. Acquiring Military Knowledge from the Encyclopedia of China. In: Proc. of the Sixth Intl. Conf. for Young Computer Scientists, Hangzhou, 2001. 368～372
19 Li Q Y. Formulae of Traditional Chinese Medicine. China Academy Press, 1998
20 Plant R T. Techniques for Knowledge Acquisition from Text. International Journal of Computer Information Systems, 1994, 35: 64～70
21 Lo H R, et al. Colored Icones of Chinese Medicine. Guangdong Science and Technology Press, 1991, 1-5
22 Lu R Q, Cao C G. Towards Knowledge Acquisition from Domain Books. In: B. Wielinga, B. Gaines, G. Schreiber and M. Vansomeren, eds. Current Trends in Knowledge Acquisition, Amsterdam, IOS Press, 1990. 289～301
23 Lu R Q, Cao C G, Chen Y H, et al. A PNLU Approach to ICAI System Generation. Science In China (Series A), 1996, 38: 1～10
24 National Committee of Codex. Codex of China (Volume 1). Chemical Engineering Press, 2000
25 Richardson S. Determining Similarity and Inferring Relations in a Lexical Knowledge Base: [PhD. Dissertation]. City University of New York, 1997
26 Richardson S, Dolan W B, Vanderwende L. MindNet: Acquiring and Structuring Semantic Information from Text. In: Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Intl. Conf. on Computational Linguistics, 1998. 1098～1102
27 Sato H, Fujimoto K. A New Approach to Semantic Word-Matching for Knowledge Acquisition from Text Containing Daily-used Words. In: M. Mohammadian, ed. Advances In Intelligent Systems: Theory And Applications, 2000, 59: 135～140
28 Schmidt G. Knowledge Acquisition from Text in a Complex Domain. In: F. Belliand F. J. Radermacher, eds. The 5th Intl. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, 1992, 604: 529～538
29 Shi D B, et al. The Medical Volume of the Encyclopedia of China. Publishing House of Encyclopedia of China, 1991
30 Tian W, Cao C G. A Framework for Extracting Knowledge of the Human Blood System from Medical Texts. In: Proc. of the 16th Intl. Conf. for Young Computer Scientists, Hangzhou, 2001. 501～505
31 Tschaitschian B, Abecker A, Schmalhofer F. Information Tuning With KARAT: Capitalizing on Existing Documents. In Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, 1997, 1319: 269～284
32 Lu R Q, Cao C G. Towards Knowledge Acquisition from Domain Books. In: B. Wielinga, B. Gaines, G. Schreiber and M. Vansomeren, eds. Current Trends in Knowledge Acquisition (Amsterdam, IOS Press, 1990) 289～301. N. Guarino, Formal Ontology and Information Systems, Proc. of the First Intl. Conf. (FOIS'98), June, Trento, Italy