汉语术语定义的结构分析和提取[1]

张艳　宗成庆　徐波

（中国科学院自动化所 模式识别国家重点实验室 北京 100080）

摘要： 本文介绍的工作是在汉语句法分析的基础上的一种应用研究，对术语的下定义问题进行了理论上的探讨。术语的定义概念提供模板和构成方式，可以作为知识发现研究的数据基础。本文针对电子学和计算机领域的语料进行了分词和词性标注处理，然后应用句法分析工具分析出句子中的短语成分，并根据汉语句子的句型结构，总结出术语定义的结构特点，自动提取定义的模板。最后根据已建立的数据和概念描述，给出了术语发现的算法。

关键词：句法分析，知识发现，术语定义

Structure Analysis and Extraction for the Definitions of Chinese Terms

ZHANG Yan, ZONG Chengqing, XU Bo

(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The work presented in this paper is a kind of application based on Chinese syntactic parsing. It is theoretic discussion to define term names. The definition of terms provides patterns and structures for term concepts and is the data basis of knowledge discovery. In this paper, the corpora of electronics and computer domain are firstly segmented and tagged with part-of-speech. Then two parsers are applied to obtain structures and phrases of sentences. According to the syntactic structures of Chinese sentences, we have summarized the structure characteristics of term definitions and automatically extracted the patterns of definitions. Finally, we describe the algorithm to define a new term according to the built data knowledge.

Keywords: Syntactic Parsing, Knowledge Discovery，Term Definition

一、 引言

术语是持有某一目的而使用的语言，分为日常生活中使用的一般性术语和科学技术专业用术语。专业术语是对一个专业概念的系统性、概括性的描述[1]。本文所做的工作主要研究专业术语的定义。术语数据库是自然语言处理领域的一个重要的知识资源。通常术语用来描述科学技术方面的概念，属于语言词汇的一部分，但又不是被人们普遍认知的词汇。术语的记录和整理与现代科学技术的发展密切相关，同时也是总结和扩展新的科学概念的需要。

术语的定义在自然语言处理领域中的研究还不是十分广泛，它为术语概念提供模板和结构方式，也可以作为特定领域知识发现研究的数据基础。目前国内还没有在术语下定义这方面进行过系统的研究。作为汉语句法分析的一个应用，如何给术语下定义有着重要的理论研究意义和应用价值。它

[1] 基金项目：国家自然科学基金资助项目（60175012）；国家 973 项目（G1998030504）

作者张艳，1973 年出生，女，北京人，博士生. 主要研究领域为自然语言处理.

不仅可以得到新的语法知识结构，还能够为自然语言处理的应用领域，如问答系统，知识发现等提供基础的知识数据库。

本文是针对电子学和计算机领域提取出相关的术语及其定义模板，通过自动获取的结构模板，给出术语下定义的实现算法。文章第二部分概述了定义的概念，并且总结了术语下定义的一些规则和方法，第三部分针对电子学和计算机领域的术语知识，自动获取了相关的定义组成结构和模板结构，第四部分是根据获取的数据，提出了术语发现的算法和结果分析，最后是结论。

二、 “定义”的概念和下定义的方法

2.1. “定义”的概念

定义就是用一个已知概念来对一个新概念作综合性的语言描述[2]。从结构上来说，一个定义可以分为被定义项（definiendum）和定义项（definiens）两部分。定义项就是对被定义项的概念性描述。它一般又可以分为属概念（genus）和种差（distinctive characteristics）两个部分。下面是定义的表示形式：

被定义项 ＝ 定义项

（被定义的概念） 属概念 ＋ 种差

下“定义”究竟是什么意思呢？它首先是把某一个概念放在另一个更广泛的概念里，而这个更广泛的概念就是最邻近的属概念，也就是“上位概念”。一个概念可以有不同层次的属概念。用属概念和种差给一个概念下定义时，究竟选择哪个层次的属概念才算合适，要看解决问题的实际需要。在一般情况下，我们总是选择与所定义的概念最邻近的属概念来下定义。因此，如何选择属概念和选择什么属概念，是术语定义的重要问题之一，是应该慎重考虑的。

种差是使被定义概念与属概念区别开来的属性，它可以是事物的性质，也可以是事物产生或形成的情况，也可以表示事物的功能、位置、形成过程等等。例如，“辐射计是测量电磁辐射量的仪器”这个定义中，“辐射计”是被定义项，“测量电磁辐射量的仪器”是定义项。定义项由属概念“仪器”和种差“测量电磁辐射的”两个部分组成。这里，种差表示“辐射仪”的用途和功能[2]。

2.2 术语下定义的方法和规则

术语下定义一般具有模式化的特点，有其独特的规则和方法，通常有三种基本类型。第一是内涵定义，就是列举被定义概念的特征来描述概念的内涵的定义，是经典意义上的定义。内涵定义中所表示出来的特征，应该是被定义概念的本质特征。第二是外延定义，是通过概念的数量范围来描述概念的外延的定义。外延定义要求一些预先的知识，这种定义假定接受定义的人事对于所列举和描述的概念已经比较熟悉。由于外延定义比较简洁明白，常常出现于法律文献中。第三叫做上下文定义，通过引证被定义概念在文章中出现的上下文来说明它的含义。因为有上下文，概念的含义往往可以被理解或猜测出来[3]。

要给一个概念下一个正确的定义，除了必须具备被定义概念所涉及的具体的科学知识之外，还必须掌握定义的规则。只有把被定义概念所涉及的具体科学知识和定义的规则结合起来，才能给概念下一个正确的定义。定义的规则通常有四条[2]：

（1）定义必须是相称的，定义项的外延和被定义项外延必须完全相同。不满足这条规则，就会“定义过宽”或“定义过窄”。“定义过宽”就是定义项大于被定义项的外延。而“定义过窄”就是定义项的外延小于被定义项的外延。因此，判断定义是否相称的标志是看定义项与被定义项是否可以互换。

（2）定义项不能直接或间接地包含被定义项。违反了这条规则，就会犯循环定义的错误。

（3）内涵定义一般不能是否定的。如果内涵定义是否定，那么就只是表示被定义项所反映的事

物不具有某种属性，而没有说明被定义项所反映的事物具有的特殊属性。这样，定义就不能揭示概念的内涵。但是，对于某些事物来说，不具有某些属性正好就是它的属性，在给反映这种事物的概念下定义时，就需要用否定的形式。

（4）定义必须使用科学的术语。为了清楚地揭示被定义概念的内涵和外延，必须用科学的术语给概念下定义，以避免对被定义概念作错误的理解。在一个定义体系中，只能使用已经定义的术语或众所周知的术语来作定义中的定义项，而不能使用含糊不清的概念，也不能使用比喻。

以上四条规则是下定义时必须严格遵守的规则，违反了其中的任何一条，都将使定义出现逻辑上的错误。

三、 限定领域术语的知识库建立

3.1 术语语料的句法分析处理

我们进行术语结构和模板的自动提取的原始语料是大百科全书[4]中的电子学和计算机领域的语料，它包含了术语的概括性定义、发展过程和实际中的主要用途等信息。把这些生语料加工成可用的熟语料需要下面的三个处理过程。第一，把原始语料处理成干净的纯文本文件，删除一些特殊符号和无用符号。例如：术语“ 安全性试验 ”的原始定义为“ 验证设备是否符合『HTK』安全防护『HT』规定的实验 ”，删除其中无用的符号 “ 『HTK』 ”和 “ 『HT』 ”。

第二，对干净的纯文本文件进行分词处理。采用的方法是 tri-gram 统计方法。由于该统计模型是基于《人民日报》的语料训练的，本文中的语料是关于电子学和计算机的专业知识，属于不同的领域范围，集外词占有相当的比例。为了消除直接分词造成的正确率不高的影响，本文采用分步的方法进行分词。

（1）用原始的 tri-gram 模型先对 20%的语料进行第一次分词，分词的正确率是 85.1%；

（2）对分完词的这部分文本进行手工校正，然后用这些分词正确的句子做训练，得到新的统计信息；

（3）用新的频率信息对剩余的语料的 20%做分词处理；

（4）循环第二、三步操作，每次都处理剩余语料的 20%，直到最后全部分词完毕。

下面通过图 1 来表示 5 次分词的结果。

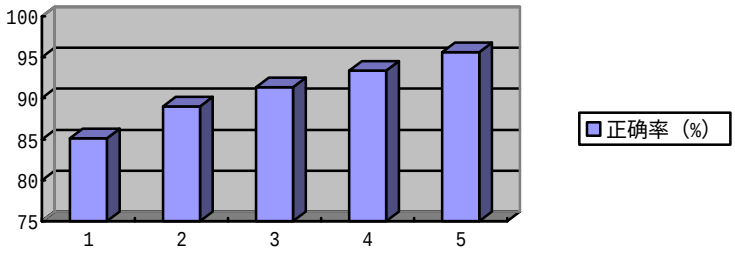

图 1 分步处理的结果比较

第三，标注经过分词的语料。上述的 tri-gram 统计模型是分词和标注一体化处理[5]，每一次分词的同时都得出相应的词性。在校正分词错误的时候，并没有把所有标注错误的词性都进行校正，只是把分错的词重新标注了正确的词性。这样，直到全部的语料用统计方法进行标注以后，最后用基于转换的规则方法获取的规则模板进行了校正。混合方法进行词性标注后的正确率为 91.20%。

经过整理后的准确、概括性的术语定义大约有 1100 个，术语的定义有其独特的地方，它的总体结构是固定的，但短语结构的子模块的组合又存在着多种形式。分析术语的定义语句结构采用了两个方法，一是在 GLR 方法[6]基础上进行改进的方法，二是线图分析方法[7]。同样的，由于分析语

料的领域差别和集外测试的数据稀疏问题，我们也用类似于分词的处理方法分步分析术语的定义。但是由于语料中术语的数量较少，我们分为两步处理，即 50%的定义语句（C_1）用原始的规则进行分析，剩余的 50%（C_2）作为集外的测试。

我们的基本的概率上下文无关文法的规则库 R_1 是从 UPenn 的中文树库中获得的，每条规则的概率反映的是该语料中的短语结构分布情况。用 R_1 中的规则分析 C_1 的定义语句，在得出初步的句法结构后，手工校正结果中的错误，然后用这部分新的标注好结构的语料重新提取规则，得到新的概率上下文无关规则库 R_2。这两个规则库各有其优点：R_1 覆盖面相对较大，R_2 更有针对性，更接近术语定义语句的分布特点。我们把 R_1 和 R_2 中的规则合并起来，但是并不是规则的频率值直接相加，对 R_1 和 R_2 选取不同的权重，得到规则库 R_3 的概率计算公式：$P(R_s) = \alpha P(R_1) + \beta P(R_2)$。通过实验确定α＝0.2，β＝0.8 的结果最好。例如表 1 中给出的规则。图 2 列出了两种分析方法随α和β的变化曲线，横坐标表示α和β的比率。

R_1 规则库	R_2 规则库	R_3 规则库
NP→NN+NN　2494	NP→NN+NN　1172	NP→NN+NN　939
NP→ADJP+NN　682	NP→ADJP+NN　60	NP→ADJP+NN　184

表 1 规则的统计

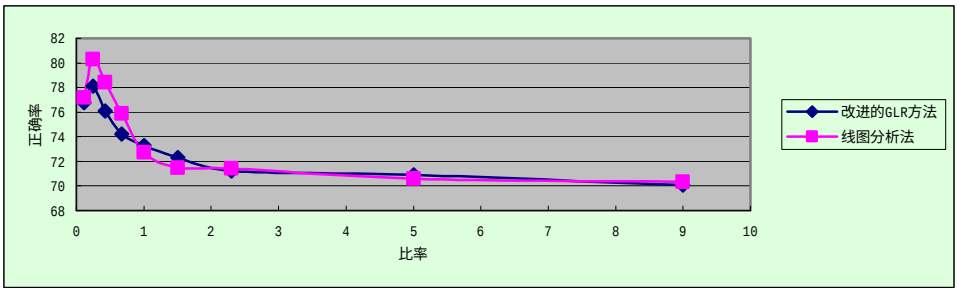

图 2　分析方法随α和β的比率变化的情况

虽然分步完成这些定义的结构分析，在一定程度上提高了短语的分析正确率，但是还存在着数据稀疏的严重问题，甚至会出现矛盾。在新的规则 R_3 的基础上，分别两种分析算法分析 C_2 中的定义语句。在表 2 中列出各部分处理的最后结果。

预处理和分析	结果正确率（%）
汉语自动分词（tri-gram）	95.64
汉语词性标注（混合方法）	91.20
汉语句法分析（1）*	78.01
汉语句法分析（2）**	80.32

表 2　预处理的结果

其中*表示改进的 GLR 算法；**表示线图分析算法

3.2 术语知识的语义词典

在术语定义中，被定义的概念与属概念有着密切的关系，属概念是被定义项的上位词。但是通常，我们把术语和属概念都看作名词，如果只从语法词性方面是很难区分开来，因此我们进一步从语义分类属性进行划分。术语和属概念的语义类以及二者之间的关系的建立主要参考《同义词词林》中的划分和分类标记。主要结构如下所列：

（一）名词（N）

（1）B 类：物统称

NBA：　统称，物，物体，器具，设备，物资，生活资料

NBB:　拟状物，网，波
NBG：　自然物
NBM：　材料
NBN：　建筑物
NBO：机器，工具，泵，机床，刀具，通信器件，

(2) D 类：抽象事物
NDA：现象，迹象，效用，反应，过程
NDB: 规律，根源，方法，技术
NDD：性质，系统，种类，结构
NDI: 事业，行业
NDK: 学说，知识，科学，学科，资料，文件，文法
NDN1：量度，温度，距离，数量

(3) C 类：地点
NCB：地点，地方，区域，线，面

(二) 动词 (V)

(1) H 类：活动
VHD：生产
VHF：交通，运输
VHI: 通信

(2) I 类：现象与状态
VIA：自然现象
VID：物体状态
VIE：传导，传播，流行，连接，纠缠，结合，割裂
VIH：变化

在术语定义的语义词典中，存放的词条主要是术语名称及其相关的信息。每个词的相关内容包括以下信息：

词条信息：术语的名称；

词条的组词信息：把术语词分词，标出中心词的语义词性；

属概念信息：标出语义词性；

属概念出现的频率：记录属概念的信息；

是否无属概念标志：主要针对解释性定义，这类定义无属概念。

术语的属概念词典是从处理后的“电子学和计算机”领域中的术语定义语料直接获得的，从分词后的定义语句找到属概念词，再根据上述的语义划分，标记其相应的语义属性。例如，术语“电子电压表”的定义为“用电子线路构成的测量电压的仪器”，它的词典信息为：

词条：#电子电压表
术语词分词：%电子＋电压表/NBO5
属概念：仪器/NBO
属概念频率值：@15
定义性质标志：无符号’$’

3.3 术语定义的模板结构

术语定义的模板是从已经标注和分析好的语句中提取的。术语的定义形式相对比较固定，规律性较强。通过分析和比较，术语定义的模板主要包括两个部分：主框架和句子的种差的语法结构。

（一）主框架

它的形式主要有两种：

1．NN1 [NN2] ::= 是|就是 + […] + 的 + NN [NP]

从左到右，符号的意思依次表示为：NN 为名词词语，[]表示可选，::=表示定义为，[…]是主要部分的结构。这个模板的意思就是术语名词 NN1，又称为 NN2 或简称为 NN2，定义为是或者就是[…]的属概念名词 NN 或名词性短语 NP。

2．NN1 [NN2] ::= 称|称为 + […] + 的 + NN [NP]

这个结构与 1 中的结构很类似，只是次序颠倒了。这个模板就把术语解释为：[…]描述的名词 NN 或短语 NP 称为术语 NN1，简称为 NN2。而这种结构的种差的起始动词往往是动词，如：研究，利用等。

（二）种差的结构

种差的结构大致分为 5 种。用下面的表 3 说明这几个部分的结构。

结构种类	具体结构	说明	分布比例
以名词（NN|NP）起始		通常是解释性定义	20%
以动词（VV|VP）起始	VP	通常以“研究、利用和使用”开头	50%
	VV NP VV [VP]		
	使 NP PP 而/CS VV		
	VP NP 用来|以便|因而（连词） VP		
以介词短语（PP）起始	PP VP NP	介词短语的语法结构主要是“在 … 中/内/之间”和“当 … 时”	20%
	PP NP		
以数量词短语起始		如，一种，一个	4%
以一些特殊的词起始	由 NP 组成|构成	起始词以介词居多，同时有少量动词，比如：由/P，对/P，以/P，把/BA，将/BA 等。	6%
	以 NP 为 NP		
	对 NP VV [NP]		
	将|把 NP VP NP，[以 VP]		
	用 NP VP		

表 3 种差的结构

在提取句子的模板的同时，又自动提取出所有的语法短语结构。这样的短语规则库中大约有 1000 条规则。其中名词短语结构（NP）有 500 条，动词短语（VP）有 300 条，介词短语（PP）有 130 条。在这些规则中，出现频率较高的有“NP => JJ + NN”、“PP => P ＋ NN ＋ LC”、“VP => VV ＋ NN”和“VP => VV ＋ VV”。当然还包括一些特殊词语构成的规则，如“VP => 被 ＋ VV”和“VP=> 受 ＋ NN ＋ VP”等等。

术语的定义绝大多数符合种差加上属概念的形式，但也有少部分的例外。这样的句子在我们的语料中大约占 10%，重要是解释术语的变化和产生的过程。而这样的术语通常是动词性的词语，比如：光电子发射、幅度测量和外延生长等；以及很难用定义的形式来描述，只能列举出所包含的事物，比如：微电子技术，其定义为“微电子技术包括系统和电路设计、器件物理、工艺技术、材料设备、自动测试，以及封装、组装等一系列专门的技术”。

四、 术语定义的自动发现算法

有了上述建立的数据知识，我们就可以从语料中自动获取术语的定义，这相似于知识发现[8]，但是由于属于特定领域的特殊应用，在算法上相对简单一些。术语下定义的过程也是从数据集中抽取和精化新的模式。主要包括数据准备和数据挖掘两个部分。我们在术语下定义时，参考了知识发现的实现算法。从整体上看包括三个部分：

1．术语名称词典查找

术语的语义词典中主要包含术语的名词，术语的分词信息和其属概念。给出一个术语名称 TermName，首先查找它的分词组成成分 SubWord（i），其中$1 \le i \le n$，从后往前看这些词，如果 SubWord（n）是名词，则它的语义类型包含于属概念 TermGenus 的语义类，那么就不再判断其他的 SubWord（i），而且定义的主框架是“ TermName 是描述…的 TermGenus ”。如果 SubWord（n）是动词，再看 SubWord（n-1），如果仍然为动词，则这个术语是解释性名称，没有属概念；如果为名词，则有两种可能，一是解释性定义，另一种为定义的主框架。因此这样的术语结构还需要进一步地判断。

例如，术语“ 电子系统工程 ”，“ 工程 ”是名词，它的属概念是“ 工程技术 ”，所以该术语的主框架一定是“ 电子系统工程是一种…的工程技术 ”。

2．术语模板结构匹配

这部分实际是判断种差的过程，即如何描述术语的特征，也是最困难的部分。对于解释性术语由于没有属概念，而它们通常又表示一个过程，所以这里统一地用“ 过程 ”来做为属概念。

描述种差仍然要用到术语的语义属性，根据术语中的分词成分 SubWord 的语义类，选择相应的动词，然后根据上述学习到种差的五个大类结构及其子类结构，最后集中到某个短语上，用提取的模板进行匹配。例如，短语“ PP ”可以用“ 当…时 ”代替。

上述过程其实是一个模板匹配的过程，而且要考虑中心词的语法语义词性。

3．反馈以及选择合适的术语定义

反馈过程是为了降低下定义过程的混淆度，通过把由算法给出的定义与原始定义相比较，根据不同之处，获取新的模板和语义搭配关系。

为了更清楚地说明术语下定义的过程[9]，我们用框图 3 来表示。

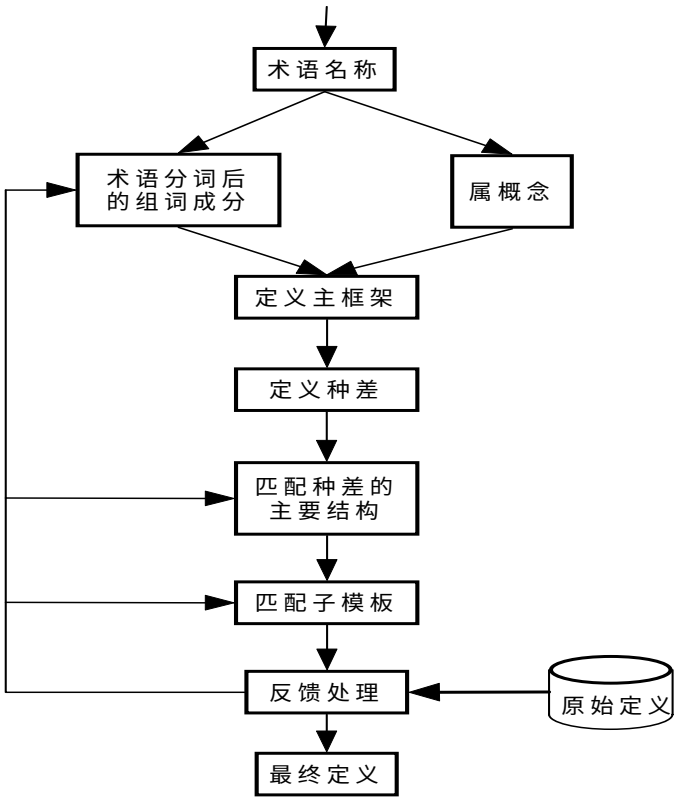

图 3 术语自动定义的算法流程

下面举个例子说明术语定义的实现过程。

根据上述的算法，我们给出一个例子说明术语的定义过程。例如，术语名称 TermName 为“ 库

存论 ”，分为下列几步：

1.查找术语名称词典，该 TermName 由 “ 库存和论 ” 两个词语组成，是名词词性，所以它不是解释性术语，因此存在着种差和属概念。从词典中查到其属概念 TermGenus 是 “ 理论、方法 ”，属于 “ NDK ” 语义类，也就是 “ 学科、理论和学说 ” 类，进而得到该 TermName 的主框架是 “ 库存论是.....理论和方法 ”。而描述种差的关联词语包括动词：研究和名词：库存、存储和问题以及连词：和。

2．差找到术语的种差结构，即如何描述属概念，这是关键的一步。从语料中统计的种差的频率值的大小来判断，首选以动词短语 VP 起始的短语结构，而且 “ NDK ” 通常用动词 “ 研究和采用 ” 作为起始词语，所以选择这个短语结构是正确的。在短语规则库中 VP 短语有 301 种，其中规则 VP → VV + NN 的概率最大，但是关联词典中有连词 “ 和 ”，短语中应该有并列成分，所以规则 VP → VV + NP 更适合，而且存在规则 NP → NN + C + NN，用已经查到的短语组合成短语，即种差可以描述成：“ 研究库存和存储问题 ”。

3．把术语的各个部分统一起来，即库存论是研究库存问题和存储问题的理论和方法。原始的定义是 “ 库存论是研究有关库存和存储问题的理论和方法 ”，所以，我们通过词语属性和句法结构获取的定义基本正确。

在术语定义的实验中，我们从《大百科全书》中的 “ 电子学和计算机 ” 领域的语料中获取了 1100 个术语名称和定义术语的句子，并且从中获取了两个主框架和 1000 个模板结构，根据所给的发现算法对这 1100 个术语重新给出定义，把获得的新定义和原始的定义进行比较。这相当于句法分析的逆过程。

例如，上一节中的术语 “ 库存论 ” 经过定义后描述成 “ 研究库存问题和存储问题的理论和方法 ”，而它的原始定义是 “ 研究有关库存和存储问题的理论和方法 ”。从意思上看，新获取的定义和原始定义基本相近，但是在句子的准确性，词语的添加、删除和排列等方面，尤其是一些非重点词和虚词，如副词、介词和连词的处理，还存在着差距，而且是一个很难解决的问题。

在评价所给定义的正确率时，需要从不同的角度考虑，例如词语之间的相似度，词语之间搭配关系，句子结构之间的相似度，而且相似度如何计算都是难题。所以我们在实验中没有给出定义的正确率。我们的实验主要是从理论上给出术语定义的结构和定义的特点，为以后的工作提供给知识基础。

五、 结论

本文主要是在汉语句法分析基础上的应用。我们的任务是如何给术语下定义。这是一个比较新的研究课题，并且国内研究得很少。本文首先描述了术语下定义的规则和方法，分布阐述了内涵定义、外延定义和上下文定义的理论方法，并且介绍了术语下定义的规则以及一些辅助手段。在这些理论知识的基础上，本文以大百科全书的电子学和计算机领域的知识为研究基础，建立了术语知识库，总结了术语下定义的句型结构和语法模板的表示形式。本文尤其是综合知识发现的理论和方法，探讨了一种用于术语下定义的自动实验算法。该算法可以在术语知识库的基础上，自动产生新的术语定义，新的科学性的概念知识，也可用于知识挖掘，编辑指导性计划的说明，问答或实现某一领域的对话系统，或辅助教学系统等，是一项有应用前景的知识处理系统。

参 考 文 献

[1] 冯志伟. 《自然语言的计算机处理》. 上海：上海外语教育出版社, 1996.10

[2] 冯志伟. 术语定义的原则和方法. 《中国术语网通讯》, 1994，第一期，pp18-.24

[3] 刘悦耕. 术语标准中的定义. 《自然科学术语研究》，1990，第二期，pp102-110.

[4] 黄鸿森. 百科全书编纂求索. 中国大百科全书出版社，1993

[5] 高山. 面向自然语言的信息智能检索[硕士论文]. 北京：中国科学院自动化所，2001

[6] Masaru Tomita. An Efficient Augment-Context-Free Parsing Algorithm. Computational Linguistics, Volume 13, Numbers 1-2, 1987

[7] 赵铁军等，机器翻译原理. 哈尔滨：哈尔滨工业大学出版社，2000

[8] 史忠植. 《知识发现》. 北京：清华大学出版社. 2002

[9] 张艳. 汉语句法分析的理论、方法的研究及其应用[博士论文]. 北京：中国科学院自动化所，2003