基于《知网》的中文信息结构抽取[1]

董强　　郝长伶　董振东
中国科学院计算机语言信息工程研究中心 北京 100083
E-mail: support@keenage.com

HowNet-Based Extraction of Chinese Message Structures

Qiang Dong　Changling Hao　　Zhendong Dong
Research Center of Computer & Language Information Engineering, CAS, Beijing, 100083
E-mail: support@keenage.com

Abstract: The Chinese message structure is composed of several Chinese fragments which may be characters words or phrases. Every message structure carries certain information. We have developed a HowNet-based extractor that can extract Chinese message structures from a real text and serves as an interactive tool for building large-scale bank of Chinese message structures. The system utilizes the HowNet Knowledge System as its basic resources. It is an integrated system of rule-based analyzer, statistics based on the examples and the analogy given by HowNet-based concept similarity calculator.
Keyword: Chinese message structure; Knowledge Database Mark-up Language (KDML); parsing; chunk;

1 引言

近年来，语块分析（chunk parsing ）或被称为浅层句法分析（shallow parsing）或部分句法分析（partial parsing）等语言处理方法成为语言技术研究的一个热点。其中印欧语言在这方面的研究已经取得了一定的成果，但是对于中文的语块分析则相对落后。其中主要的原因在于：相对于印欧语言，中文没有那么丰富的形态变化，中文的词类与句法功能不是一一对应的，中文的词、短语、句子之间的界限是模糊的。除此而外，甚至可能是更重要的原因在于：中文的结构更加依赖于语义的制约。

本文提出中文信息结构这样的概念，并据此构建了基于知网的中文信息结构抽取器。该抽取器从中文的语义出发，以知网知识系统作为其基础资源。中文信息结构抽取器的功能主要包括两个方面：（1）构建、管理和维护《知网-中文信息结构库》的数据；（2）是分析并抽取真实文本中的中文信息结构。其技术关键是：第一，它对于中文的词、词组、短语进行一体化的处理；第二，它主要是基于语义的。第三，它是基于规则匹配和基于大规模实例库及相似度比较的结合。中文信息结构抽取器利用已有的中文信息结构的模式与实例编写规则，再利用这些规则到真实的文本中抽取所需的中文信息结构来构建更大规模的中文信息结构库。

2 中文信息结构

信息结构（message structure）是由两个或两个以上的字、词或短语构成，句法和语义合理，并传达了特定信息的结构。该结构内不含有介词、助词、连词、标点。该结构内部允许呈递归形态。信息结构是信息理解（message understanding）的基础。信息结构是依赖于特定语言的，不同的语言可以表达相同的信息，但有着不同的信息结构。试以"通缉犯"这样一个词语为例。其中有两个基本单元“通缉”和“犯”，“通缉”为动词，“犯”为名词，因此从句法结构看是 V＋N 的定中结构；从语

[1]基金项目：国家自然科学基金项目 60372064；国家语言文字应用“十五”项目 YB105-50

义关系结构看，前者是“行为动作”，后者是“受事”；而信息结构不仅能反映“行为动作”和“受事”之间的关系，而且还能够反映出是“这是一个人，这个人是被通缉的对象，他是一个有罪的人（“有罪”这一信息是含在“通缉”这一词语中的）”。
中文信息结构是中文中句法和语义合理的一个语言片段，它可能是传统被认定的词语，也可能是一个比词语更大的语言片段。中文信息结构的描述对象是：由中文词语所表述的概念。《知网》规定了最基本的运算单元是：万物、部件、属性、属性值、事件、时间和空间等 7 大类。

(a) 万物
平民-百姓，车-辆，运动员-们，桌-椅，赵-大夫，杨-队，大学-老师，海外-房地产-投资-公司，电影-演-员，采访-者，外交-界，物理学-家，食蚁-兽，手术-室，美食-城，西药-房，设备-保护-装置，洗涤-剂，豆瓣-酱，长-袖-衫，大-黄-狗，锦绣-河山，金-光，酸-梅，业余-棋手，一-本-杂志

(b) 部件
头-顶，脸-部，腹-腔，科-室，手-套，熊-腰，文章-段落，牙-根，树-枝，杏-仁，山-顶，句-首，屋-脊，桌-面，车-身，笔-套，学校-各处室，鸡-腿，尖-下颌

(c) 属性
浓-淡，安-危，高-度，通畅-性，员工-人数，药-效，地球-籍，温-差，氧气-消耗-量，出生-率，社会-环境，警-民-关系，主要-特点，本届-会议

(d) 属性值
朱-红，没-水平，高-性能，双-色，可-悲，防-洪，耐-寒，翠-绿，中-高级，银-灰，深-蓝，更加-重要，功能-正常，不-称职，干-干-净-净，快-起来，软-下去，十-七，第-五，三-分之-二，百分之-十

(e) 事件
安排-工作，卖-书，参观-学校，供应-部队，植-树，订-计划，新闻-报道，物价-检查，药品-生产，拜-寿，爱好-体育，爱-看书，称-重，恢复-军籍，度过-难关，中断-比赛，禁止-吸烟，爱-漂亮，睡-三小时，打-牌，非常-爱护，两-年-写-四-本，深-感，怒-斥，盗-墓

(f) 时间
道光-三十-年，公元前-368-年，四-月，六-号，第五-天，六-点，1939 年-9 月，1 月-24 日，今天-凌晨，上午-九时，12 月 9 日-星期四，八时-三十分，明天下午-五点，四个星期-内，本世纪-末期，上午八时-整

(g) 空间
江西-省，加拿大-多伦多，辽宁-本溪，厦门-湖里区，北京-广渠门大街，香颐路-宁宜小区，秀水园-1-号，6 号楼-5-号，中山北路-三-段，丽都-饭店，中山-公园，闪光-点，战争-地点，网-上，古-国

现在有很多人在做中文命名实体辨识（**Chinese Named-entity Recognition**）的工作，在中文命名实体辨识中的主要是抽取人名、地名、组织机构名、时间、数量等**[4][5]**，而中文信息结构不仅涵盖了这些内容，并且其涉及的范围较目前流行的中文语块更宽泛。我们研究中文信息结构的出发点是：既然我们已经认定了世界上只有 **7** 大类概念，我们将通过对于中文信息结构的研究来发现中文是如何表达或描述这些概念的。下面我们就“万物”这个类别，在展开一点加以说明。中文在描述

“万物”时，都采用了哪些结构呢？试看下面的例子。

（a）描述“万物”作为成员角色及其整体的关系

其中的一个结构（简化）是：(组织/场所) [来源整体] <-- (人,职位)，如：

公司-总经理，图书馆-馆员，小卖店-老板娘，社科院-院长，内科-主任

（b）描述“万物” 作为被领属物

其中的一个结构（简化）是：(地方,专) [限定] <-- (组织)

美国-国会，美-军，中国-人大，俄-杜马，台湾-情治单位，以色列-摩萨德，德国-央行

（c）描述“万物”作为施事角色及其行动的关系

其中的一个结构（简化）是：(事件,行动) <-- [施事] (人/拟人)，如：

公诉-人，捕蛇-人，侦察-兵，救生-员，采购-员，收货-人，来京务工-人员，辍学-生

（d）描述“万物”作为被支配角色（如受事等）及其行动的关系

其中的一个结构（简化）是：(事件,行动) <-- [受事/成品受事/内容/对象/领属物] (万物)，如：

雇-员，展-品，弃-婴，保护-区，在押-犯，进口-货，废弃-机场，参考-资料，处理-对象

（e）描述“万物”作为销售场所及其所销售物品的的关系

其中的一个结构（简化）是：{(物质) [领属物] <-- <事件,行动,买>} <-- [处所] (组织/场所)，如：

书-店，鞋-店，银-楼鞋帽-店，中药-店，西药-店，西饼-店，床上用品-商店，自行车-行

（f）描述“万物”作为成品及其材料的的关系

其中的一个结构（简化）是：(材料/无生物) [材料] <-- (人工物)，如：

蔗-糖，草-席，玻璃-杯，葡萄-酒，塑钢-门窗，柏油-马路，汉白玉-栏杆，皮蛋瘦肉-粥，

（g）描述“万物”作为成品及其材料的的关系

其中的一个结构（简化）是：{(材料/无生物) [材料] <-- (制造/辫编)} <-- [成品受事] (人工物)，如：

铝-制-品，毛-织-品，全棉-织-品，豆-制-品，玻璃-制-品，纯羊毛-织-物，

（h）描述“万物”作为销售场所及其所销售物品的的关系

其中的一个结构（简化）是：{(属性值)/（数量值） [修饰] <-- (部件)} <-- [整体] (物质)，如：金-发-女郎，白-胡子-老头，圆-领-衫，高-领-衫，长-把儿-铁锹，独-臂-英雄，双-缸-洗衣机，

3 中文信息结构抽取器的规则与模块

根据我们多年建设知网和中文信息结构库的实践，我们可以肯定中文词语间的组合主要基于语义。在我们的中文信息结构库 2000 版中，中文信息结构（基于语义的）有 271 个，但与之对应的句法结构仅 58 个。下面的结构在句法上都是 N+V 的结构，但它们的语义结构是很不同的：货物-运输（受事），铁路-运输（手段），汽车-运输（工具），春节-运输（时间），海洋-运输（处所），人们-运输（施事）等。如果仅仅将它们捆绑在一起（如树库所为）或者仅仅把它们分类为主谓、定中等结构（如传统语言学的语法所为），对于信息处理都是不够的。因此，中文信息结构抽取器主要采用了基于语义的方法。我们在中文信息结构抽取器中构建了一系列的语义规则与模块。图 1 所显示的是我们的抽取器工作的一个结果，即根据一条有关“时间结构”的语义的规则，来抽取“当天 上午 十 点”

这一类型的信息结构。

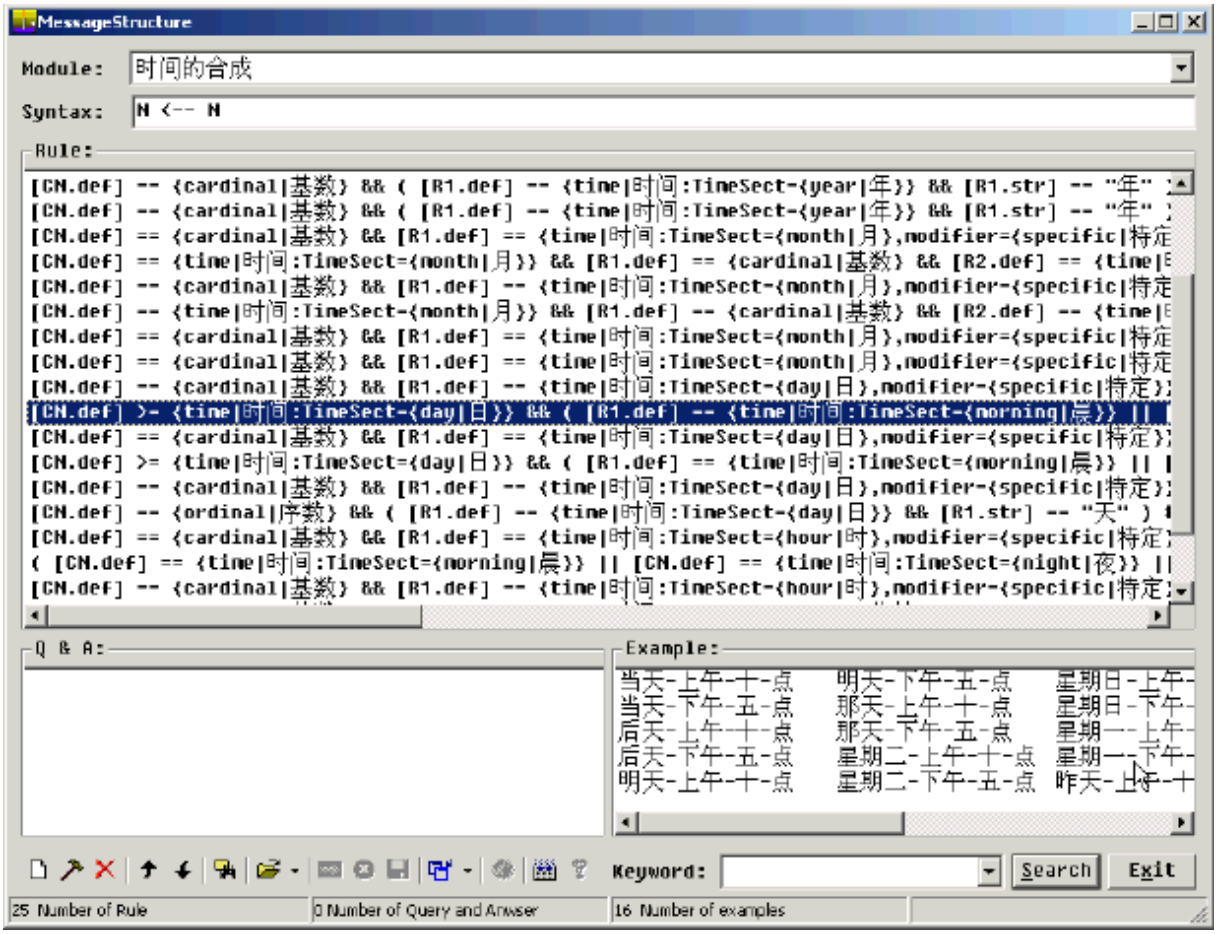

图 1：中文信息结构抽取器概况

3.1 规则与模块的形成

在中文信息结构抽取器中，我们构建了一系列的规则与模块来实现对中文信息结构的抽取。其中的规则是以《知网-中文信息结构库》为原始资料进行构建的。这是因为《知网－中文信息结构库》中的中文信息结构是从大规模真实语料中抽取，并经人工选择，确定取舍的，其中共总结了 **271** 种中文信息结构，附带着一万多实例，总字数十万余。虽然就其规模而言它还只能算是一个雏形，但就其所包含的模式而言应该说已趋于成熟。

3.2 中文信息结构抽取器中的模块

我们根据规则所能够实现的功能，将规则分配到不同的模块，如时间模块用来处理时间类中文信息结构，机构名模块则用来处理表示各种机构的中文信息结构。每一个模块下都有若干规则，并且这些规则会按照不同的优先级进行排列，将优先级高的规则放在前面，这样当调用到这个模块时，就会先调用优先级高的规则。而规则的优先级又是按照规则规定的严格与否划分的，规定严格的规则，它的优先级就高。因此优先级的高低确定了同一个模块中规则的调用顺序。

3.3 中文信息结构抽取器中的规则

每一条规则都是由规则的条件和规则的结论两个部分构成。对于规则条件的描述采用了知网知识系统描述语言（**KDML**）的描述方法。规则的条件从信息结构的每个组成部分的意义出发,规定了具体的

中文信息结构的各个部分之间的语义关系。规则的结论发出一种动作，如：开始扫描文本，调用某个规则模块，捆绑一个中文信息结构等等。例如下面的规则：

规则的条件：**[CN.def]**① **== {cardinal|基数} && [R1.def.class]**② **== {ActUnit|动量}**

规则的结论：**@chunk(R1,CN).**

例子：三-下;两-趟;

（注①：**[CN.def]**表示当前扫描节点的概念定义。）

（注②：**[R1.def.class]**表示当前扫描节点的下一个节点的概念定义的类。）

在规则的条件部分，我们首先规定了当前的节点的概念定义为**{cardinal|基数}**，接着规定当前节点的下一个节点的概念定义的类为**{ActUnit|动量}**，**{cardinal|基数}**和**{ActUnit|动量}**都是在知网中定义的义原。我们看其中“三-下”这个例子，它符合这个规则所规定的条件，由此也就确定了“三-下”这个例子中“三”和“下”这两个部分的意义。其中“三”在知网中只有一个意义，是：**DEF={cardinal|基数}**。而“下”在知网中共有 **20** 个意义，而按照这条规则的规定，它的概念定义的类应该是**{ActUnit|动量}**，因此我们就确定了这里的“下”的意义为：**DEF={ActUnit|动量:host={event|事件}}**。当一个信息结构符合了某一条规则的时候，这个信息结构中每个部分的意义都被确定了。

这个规则的结论是以 **R1** 节点为中心进行捆绑，因此就把“三-下”中的“三”作为“下”的子节点捆绑了，那么“三-下”这个信息结构的意义就被抽象成“下”的意义，即 **DEF={ActUnit|动量:host={event|事件}}**。值得指出的是“三”的意义并没有被丢掉，而是仍然被保存的，以便于进一步捆绑以及意义的辨识时的使用。

下面我们来看“爱-生病”和“爱-游泳”这样两个信息结构，从句法结构上看都是“**V**＋**V**”的结构，但是从他们的语义结构来看，前者是**(**事件**,**状态**,**易于**) --> [**内容**] (**事件**)**，后者是**(**事件**,**状态**,**情绪**/**态度**/**意向**/**感知状态**) --> [**对象**] (**事件**)**，即具有不同的语义结构，这时，我们就会将这样的两个信息结构编写不同的规则，如下：

“爱-生病” [CN.def] == {AptTo|易于} && [R1.def.class.root] == {event|事件}

“爱-游泳” [CN.def] == {FondOf|喜欢} && [R1.def.class.root] == {event|事件}

对于规则的描述采用了知网知识系统描述语言（**KDML**）的描述方法**,**这是因为**:**中文信息结构体现了概念与概念之间以及概念的属性与属性之间的关系，这也是它与普通意义上的中文语块之间的区别。在知网知识系统中，共定义了 **2215** 个义原以及 **90** 个动态角色，通过 **KDML** 的描述，这些义原与动态角色得以紧密联系在一起。中文信息结构恰恰需要这样一种描述语言来体现各种信息结构中概念之间以及概念的属性与属性之间的关系。另外，知网知识系统的知识库是利用知网知识系统描述语言（**KDML**）对各种语言的词语所代表的概念进行描述的，它是一个面向计算机的可以计算的语义资源。利用 **KDML** 对中文信息结构抽取器中的规则进行描述使中文信息结构抽取器与知网知识系统结合起来。从理论上讲，这使得知网的知识体系在中文研究中得到了进一步的延伸。从应用的角度来看，在中文信息结构的抽取过程中，通过知网知识库进行组词，使被切分的每一部分的意义都变得能够计算，同时 **KDML** 使得每一种信息结构也都成为可以计算的对象。这样，规则中所规定的某一部分应该具有的语义与被切分的每一个部分之间就可以放在一起进行意义的计算了。

就上面的例子有人会问：“如果文本中是‘爱发脾气’，抽取器又如何能知道选择哪一条规则？”这正是我们的策略与传统的基于规则的系统的不同。在《引言》中我们指出我们的任务之一是建设大规模中文信息结构库。在结构库里有大量的实例，以便我们引入统计概率的方法，同时还要特别介绍，我们有已经完成的概念相似度计算器，例如就以“发脾气”为例，根据我们的概念相似度计算器的计算，“发脾气”与“生病”的相似度为 **0.018605**；而与“游泳”的相似度为 **0.009639**。这样

也可以帮助我们判别歧义。

3.4 规则控制程序的策略

在中文信息结构抽取器中采用了规则控制程序这样一种策略。这种策略使得语言工作者与计算机工作者的工作相互分离，最大限度的发挥了各自的长处。对规则和模块进行的填加、修改、删除，制定规则的优先级，什么时候调用什么样的模块等等这些工作都可以由语言工作者单独完成，因为语言工作者最了解各种语言现象。计算机工作者只需要完成对这些规则与模块的解析。对于规则与模块的解析是通过信息结构解析器完成的。

3.5 信息结构解析器

如果不能对上述的规则与模块进行解析，那么中文信息结构抽取器就不能被激活，对真实的中文文本进行信息结构的抽取也就无从谈起。于是，我们构造了信息结构解析器。信息结构解析器是由计算机工作者完成的，用于对规则中所使用的各种描述符号、关系符号进行解释，并对其中的 **KDML** 语言进行解析，从而完成规则与信息结构之间语义的匹配。规则与信息结构语义之间的关系匹配是信息结构解析器最为重要的部分。

3.6 规则举例

文本：去年冬天我去了两趟哈尔滨。

中文信息结构的抽取结果如图 **2**，其中共有两个中文信息结构，分别是：“去年**-**冬天”和“两**-**趟”。

其中中文信息结构“去年-冬天”抽取结果分析如下：

中文信息结构：去年冬天

组合情况：去年-冬天

适用规则：[CN.def] >= {time|时间:TimeSect={year|年}} && [R1.def] == {time|时间:TimeSect={winter|冬}} # @chunk(CN,R1).

中文信息结构“两**-**趟”抽取结果分析如下：

中文信息结构：两趟

组合情况：两-趟

适用规则：[CN.def] == {cardinal|基数} && [R1.def.class] == {ActUnit|动量} # @chunk(R1,CN).

与其他的中文语块抽取系统不同的是，经过中文信息结构抽取器处理后的中文信息结构内部的语义关系是清晰明确的，其每一个部分的语义都被确定并保留了。这样一来就给我们的研究带来一个新可能，即我们可以针对已被抽取出来的信息结构，进行问与答。例如针对第 **2** 节的（**a**）我们可以有如下的问与答：

（**a**）描述“万物”作为成员角色及其整体的关系

其中的一个结构（简化）是：N1(组织/场所) [来源整体] <--N2 (人,职位)

Query1: 谁？**/** 什么人？

Answer1: N1 + N2

Query2: 他（她）是做（干）什么的？**/** 他（她）的职务是

Answer2: N1 + N2

Query3: 那是哪儿的 **N2**？

Answer3: N1“的”

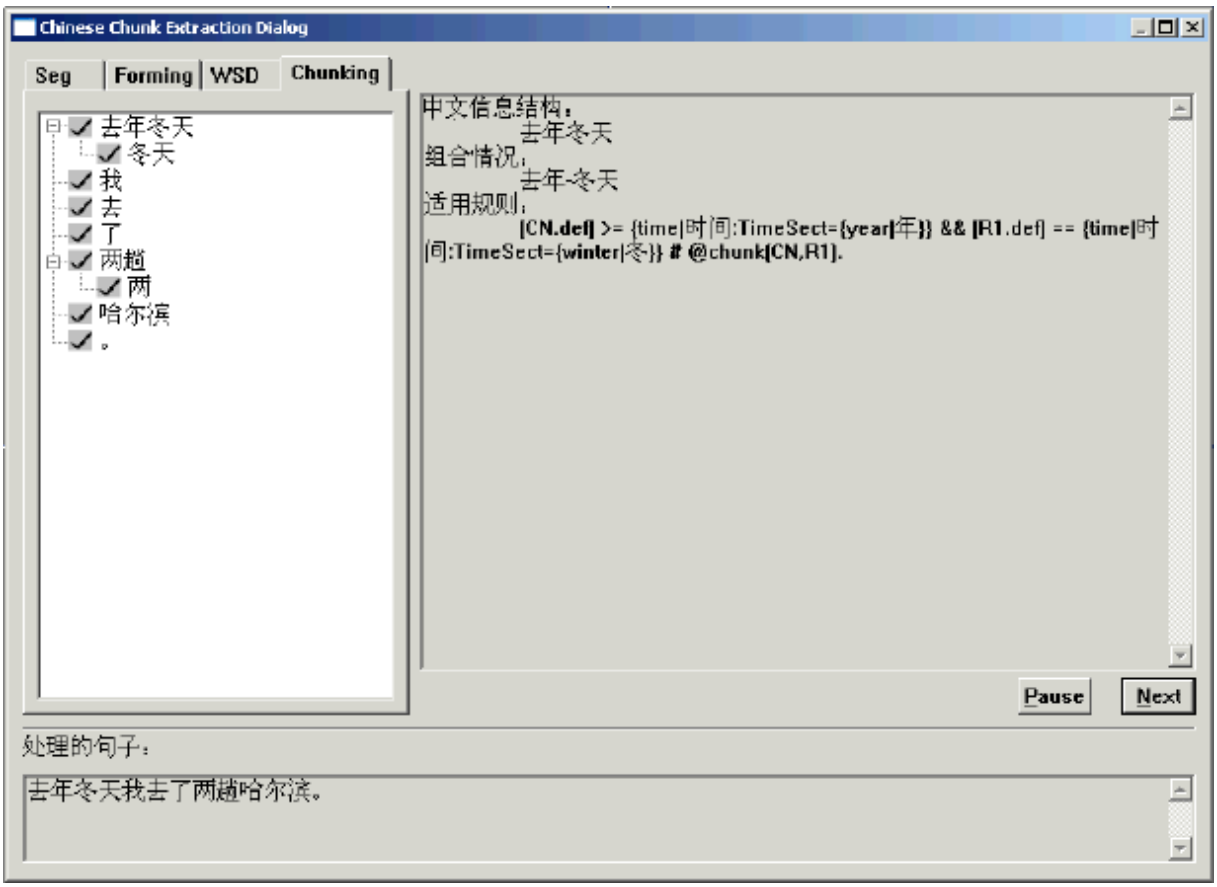

图 2：抽取结果示例

4. 结论

知网人一直致力于面向计算机的中文信息处理，因此，由知网人开发的中文信息结构抽取器也必然是面向计算机的系统。中文信息结构抽取器充分的利用了知网知识系统这一语义资源，实现了对中文信息结构的自动抽取，成为建设知网－中文信息结构库的必不可少的工具。中文信息结构抽取器目前基本完成了对时间、数量、地名、重叠词等的自动辨识，在下一步的研究中，我们将进一步完善已有的模块以及构建针对更为复杂的中文信息结构的自动辨识模块。

参考文献

［1］ 董强，郝长伶，董振东，基于《知网》的中文语块抽取器，语言计算与基于内容的文本处理（全国第七届计算语言学联合学术会议论文集），孙茂松、陈群秀主编，清华大学出版社

［2］ 颜国伟、谭慧敏，基于《知网》的常识知识标注，中文计算语言学期刊，第 4 卷第 2 期，1999

［3］ 汉语计量与计算研究，邹家彦，香港城市大学语言资讯科学研究中心，1998

［4］ Jorn Veenstra, Memory-based text chunking, In Nikos Fakotakis Machine Learning in human language technology, workshop at ACAL99, 1999

［5］ Jian Sun, Ming Zhou, Jianfeng Gao, “A Class-based Language Model Approach to Chinese Named Entity Identification”, *Computational Linguistics and Chinese Language Processing,Vol.8, No.2, August 2003, pp.1-28*

［6］ Hua-Ping ZHANG, Qun LIU, Hong-Kui YU, Xue-Qi CHENG, Shuo BAI, “Chinese Named Entity Recognition Using Role Model”, *Computational Linguistics and Chinese Language Processing, Vol.8, No.2, August 2003, pp.29-60*

网络资源：http://www.keenage.com