基于规则方法的汉语到语义网络语言的转换研究

张旭洁[1]，夏幼明[1]，刘冠晓[1]，宋亚林[1]

（1.云南师范大学计算机科学与信息技术学院，昆明，650092）

摘　要： 让计算机自动获取知识是智能系统研究者一直以来的梦想，而作为人机交互接口的自然语言理解、知识表示、知识获取的研究是目前非常热门的领域。语义网络语言是一种具有较强表示能力的知识表示方式[1]。用语义网络语言构建的知识库有助于运用语言结构进行推理，回答查询[2]。本文从依存关系树库中提取出依存三元组到语义网络语言的转换规则，并使用XML 标记语言实现了对转换规则的管理，包括规则的添加、删除、修改。在依存句法分析的基础上，运用从依存关系树库中提取的规则，找到了依存三元组与语义网络三元组表示语言的一一对应转换关系，从而实现了由汉语到语义网络语言的转换。

关键词： 知识获取；依存关系；规则；语义网络语言

Study on Rule-based method for the Transformation of Chinese and Semantic Network Language

ZHANG Xu-jie[1]，XIA You-ming[1],LIU Guang-xiao[1],SONG Ya-lin[1]

(1.Department of Computer Science and Information, Yunnan Normal University, Kunming,650092)

Abstract: It is researcher's dream of the intelligence system to let the computer obtain knowledge automatically, and it is a very hot field to study at nature language understanding, knowledge representation and knowledge acquisition. Semantic network language is a kind of language which has stronger ability to represent knowledge. The repository which constructed of semantic network language is helpful to using the structure of language to reasoning and answering questions. We get the transformation rules from dependency relationship tree-bank and use XML managing the set of rules, including these operations: accession, deletion, modification. Based on the dependency parsing and the transform rules we have realized the transformation of these two languages.

Keywords: Knowledge Acquisition; Dependency Relation; Rules; Semantic Network Language

1　研究背景

知识获取的基本任务是对专家系统获取知识，建立起健全、完善、有效的知识库，以满足求解领域问题的需要。目前知识的获取主要是由工程师与专家系统中的知识获取机构共同完成的。知识工程师负责从领域专家那里抽取知识，并用适当的模式把知识表示出来，而专家系统中的知识获取机构负责把知识转换为计算机可存储的内部形式，然后把它存入知识库。在存储过程中，要对知识进行一致性、完整性的检测[3]。

知识库知识的获取主要有：非自动知识获取、自动知识获取、半自动知识获取，三种方法[3]。

在基于语义网络语言（Semantic Network Language SNetL）的知识库构建中，由于直接手工录入知识速度比

基金资助：非规范知识处理的基础理论及处理技术研究（项目编号：04F00062）

作者简介：张旭洁（1980－），女，助教，硕士，E-mail:zhangxujie@gmail.com

较慢、容易出错，并且不同的人对同一语句描述的形式不同，缺少转换标准与规则，致使在实现智能问题求解的过程中无法有效或正确的利用这些知识。为了提高手工知识转换的速度，并制定相同语法结构的自然语言到语义网络语言的转换规则，本文构建了一个由汉语到语义网络知识表示语言转换的半自动知识获取系统，如图 1。

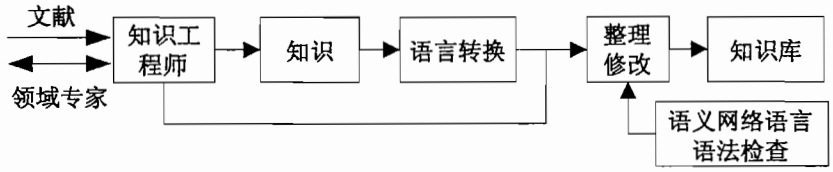

图 1 半自动知识获取

Figure 1 Semi-automatic Knowledge Acquisition

2 语言转换系统框架

2.1 语言资源和语言分析工具

在基于规则的汉语到语义网络语言的转换系统中采用了以下语言资源和语言分析工具：

（1）中文依存关系树库：使用了哈工大信息检索实验室提供的依存树库

（2）汉语分词一词性标注工具：使用了哈工大信息检索研究室提供的汉语分词模块

（3）汉语依存句法分析工具：使用了哈工大信息检索研究室提供的依存句法分析模块

（4）语义网络语言语法检查模块

2.2 系统框架

汉语到语义网络语言的转换系统框架，如图 2。

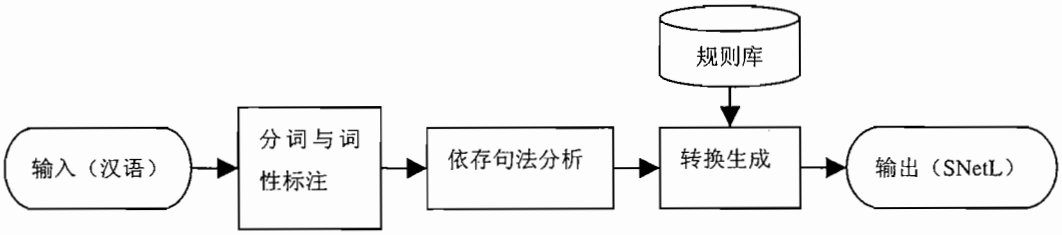

图 2 语言转换框架

Figure 2 Language Translation

输入汉语语句通过哈工大信息检索实验室提供的分词与词性标注还有依存句法分析模块的分析后生成依存三元组。下面给出依存三元组的定义：

定义 2.1（**主成分**）依存关系内，依存关系的发出者称为主成分。

定义 2.2（**支配成分**）主成分以某种依存关系支配其从属的成分，称为支配成分。

定义 2.3（**依存三元组**）结构为：主成分_支配成分（依存关系）的三元组称为依存三元组。

例 1：句子："信息是现实世界的反映。"经过依存分析后得到的分词结果与依存三元组表示如下。

分词结果：[1]信息/n [2]是/vx [3]现实/n [4]世界/n [5]的/ue [6]反映/vn [7]。/wp [8]<EOS>/<EOS>

依存分析结果（依存三元组）：[4]世界_[3]现实(ATT)；[2]是_[1]信息(SBV)；[5]的_[4]世界(DE)；[6]反映_[5]的(ATT)；[2]是_[6]反映(VOB)；[8]<EOS>_[2]是(HED)

其中每个单词前面的"[数字]"记录了单词号，即单词在整个句子中为第几个单词；"/字母"表示词性，<EOS>表示依存关系树的根节点。在哈工大信息检索研究室提供的依存句法分析模块中定义了 24 个依存关系，例如：HED（核心）、SBV（主谓关系）、VOB（动宾关系）、ATT（定中关系）、ADV（状中关系）等等。

语义网络由美国心理学家奎廉（M. R. Quilian）于 1968 年在研究人类联想记忆时提出。它通过概念及其语义关系组成的有向图来表达知识、描述语义。一个语义网络是一些以有向图表示的三元组（节点 1，弧，节点 2）

连接而成的。其中节点表示概念，弧是有方向的，指明所连接节点的语义关系[4]。

语义网络语言的一般形式表示为：

（节点 1：标号，关系：标号，节点 2：标号）；

其中，标号对节点和关系做进一步说明，其作用相当于自然语言中的定语和状语。

下面给出的例子说明语义网络语言的使用。

例 2：用 SNetL 描述如下的事实：阿布是“非典”疑似病人。

（阿布，是：L1，病人：L2）；

L1：（是，degree，5）；

L2：（病人，类型，非典）；

第一行称为 SNetL 主关系式，第二、第三行称为 SNetL 标号关系式，无论是主关系式还是标号关系式都称为 SNetL 关系式。

汉语到语义网络语言的转换，实质上是依存三元组到语义网络语言三元组的转换。

3 汉语到 SNetL 的转换

3.1 依存三元组到 SNetL 的转换思想

要把依存三元组转换为 SNetL 三元组就必须找到它们之间的对应关系。

在把一个自然语句转变成一个 SNetL 三元组的时候，总是从句子的主谓或主谓宾成分入手。通常情况下主语总是动作的发出者，通过谓语动词，作用到宾语。句子中的主语作为 SNetL 三元组主关系式的节点 1，宾语作为节点 2，而谓语作为关系，也就是说主语与宾语这两个概念通过谓语动词这一关系进行联系，构成 SNetL 的主关系式。例如：“小猫喝牛奶。”，就可以直接转换为 SNetL 三元组：（小猫，喝，牛奶）；而利用依存分析结果我们就可以直接获得一个句子的主、谓、宾成分，构成 SNetL 的主关系式。

再进一步，句子的主语，谓语，宾语都是有修饰成分的，例如修饰主语的定语，修饰谓语的状语等等，这些成分在手工生成 SNetL 三元组时是以标号关系式的形式出现的，用以进一步说明某个概念的性质、属性等等信息，例如：“白色的小猫喝牛奶。”就可以转换为 SNetL 三元组：（小猫：L1，喝，牛奶）；L1：（小猫，颜色，白色）；我们同样可以利用依存关系来确定修饰句子主要成分的内容，并转换为 SNetL 的标号关系式。

规则的提取，是为了从依存三元组中的依存关系找到把这些句子成分转换为 SNetL 三元组的对应关系。

3.2 依存三元组到语义网络语言转换规则的提取与管理

从依存关系三元组到 SNetL 三元组的转换需要了解依存关系三元组中的依存关系，根据依存关系确定此依存三元组中的主成分与支配成分在 SNetL 三元组中充当什么样的成分；即判断这些成分是作为 SNetL 主关系式中的节点、关系；还是标号关系式中的节点、关系，这些都需要对实际运用中的汉语自然语句进行分析与总结才能提取。

定义 4.1：转换规则是形如“规则类型＋规则条件→规则动作”的表达式。

（1）规则类型：规则的类型由依存关系、主成分词性、支配成分词性三个信息来确定。程序对哈工大中文依存关系树库进行提取，共获得 1213 种类型。通过计算，得到了组成不同依存关系的词性组合占此语料中这类依存关系总数的比例。

（2）规则条件：规则条件，则由预处理确定，用来判定依存三元组中的主成分与支配成分是否作为 SNetL 中的主关系式中的节点或关系。

（3）规则动作：有确定的规则类型和确定的规则条件就可以确定唯一的转换操作。在依存三元组到 SNetL 三元组的转换中对应的转换操作基本类型有三种：第一是省略，即此依存三元组中的主成分与支配成分都不再运用到 SNetL 中；第二是成词，即按照某一顺序合并依存三元组中的主成分与支配成分，使它构成 SNetL 关系式中的一个节点、关系；第三就是成句，即依存关系三元组中的主成分与支配成分按照某一组合顺序构成 SNetL 中的一个关系式。通过对语料的手工分析，系统中定义了 18 中规则动作，分别对以上三种基本操作进行了细化

处理。

（4）规则管理：程序分析获得的 1213 个规则类型决定了在依存三元组到 SNetL 关系式的转换中就会遇到这么多中的判断情况，如果把这些规则都以条件语句的判断形式编入程序的话，程序的可读性还有修改维护程序就成了一个及其耗时费事的工作。

系统运用 XML 构建了规则树，如图 3，只需要获得规则类型和规则条件，就可以从树中获得对应的规则动作。规则树的组成：第一层根节点、第二层依存关系、第三层词性组合、第四层规则条件、第五层规则动作。整个规则树以 XML 文档的形式保存，可以方便的对树中各层内容进行修改、添加、删除操作，并可以根据实际情况修改规则动作，图 3 给出了规则树编辑框图，实现了对规则集的管理。

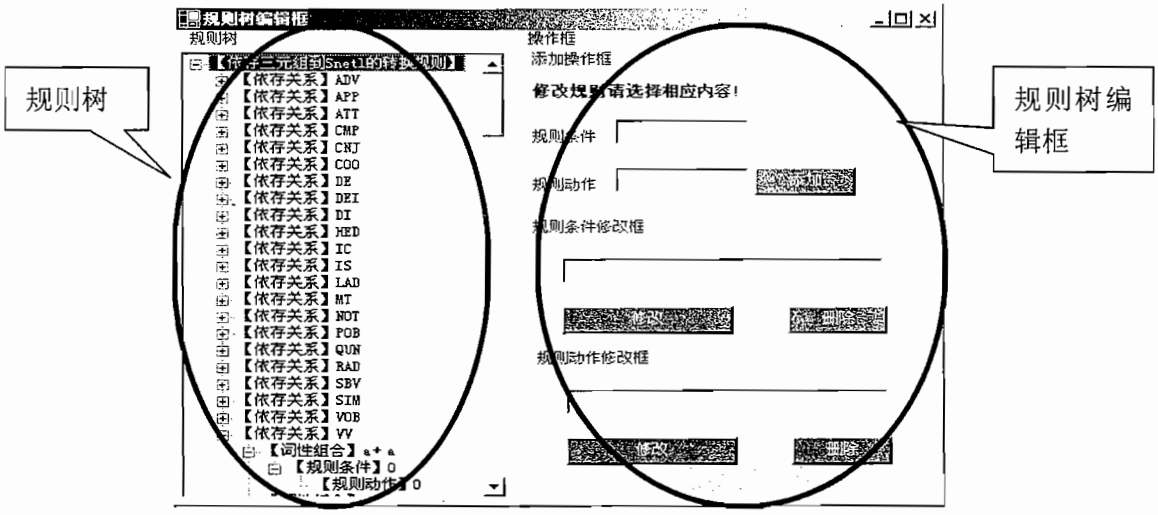

图 3 规则树与规则编辑框

Figure 3 Rule tree and Rule editor box

（5）默认规则：从 1 万句语料中提取的规则是不能满足实际情况的。系统中使用出现频率最高的依存关系的转换规则为默认规则，默认规则由依存关系与规则条件来确定。默认规则的使用，提高了程序的鲁棒性。

3.3 依存三元组到 SNetL 的转换算法

算法名称：依存三元组到 SNetL 语义网络三元组转换算法

输入：哈工大依存分析结果（分词与词性标记集、依存三元组集合）

输出：SNetL 语义网络三元组

算法过程描述：

过程 1：预处理确定规则条件

首先找到句子的 HED（核心）依存关系，其支配成分就是句子的中心动词。然后搜索以句子中心动词为主成分的依存三元组，进行规则条件的标记，完成转换前的预处理。

过程 2：通过依存关系、词性组合、规则条件从规则树中获得规则动作

规则动作为 0，表示此规则动作需要添加或者是规则集中没有此规则，程序运行中将调用默认规则处理函数，获得默认的处理规则动作。同时也可以根据需要调用规则树编辑框，对规则集进行管理维护。

执行这些规则动作生成的 SNetL 三元组标记为“句子”、“主谓句”、“动宾句”、“主句”。其中表示 SNetL 主关系式的有“主谓句”、“动宾句”、“主句”，结构分别为：“(主，谓，)；”、“(，谓，宾)；”、“(节点 1，关系，节点 2)；”，其中“主句”表示此三元组是 SNetL 的一个主关系式。这些三元组都记录在 SNetL 临时队列中，用以生成顺序正确的 SNetL 语句。

过程 3：调整这些没有顺序的 SNetL 三元组，使这些三元组构成标号索引正确的 SNetL 语句

通过句子类型为“主谓句”、“动宾句”、“主句”的 SNetL 三元组生成 SNetL 的主关系式，并运用堆栈结构实现深度优先搜索修饰节点和关系的标号关系式的过程；SNetL 主关系式和修饰其节点与关系的标号的关系式依次放入 SNetL 队列中。当搜索完所有**过程 2** 构造的主关系式节点与关系的修饰关系式后，程序结束。此时 SNetL

队列中记录的就是经过转换与顺序化的 SNetL 关系式。

以句子："信息是现实世界的反映。"为例。经过第一步与第二步的处理以后，SNetL 临时队列中存放的 SNetL 三元组如下：

L1:(现实世界，修饰，反映);【句子】

(，是，反映);【动宾句】

(信息，是，);【主谓句】

第三步，程序首先生成的主关系式是："(信息, 是, 反映);" 然后在 SNetL 临时队列中搜索修饰其节点与关系的标号关系式；然后再构造其它主关系式，直到不能再构造新的主关系式程序终止。SNetL 队列中记录了最后生成的 SNetL 语句，如下：

(信息, 是, 反映:L1)；

L1:(现实世界, 修饰, 反映)；

4 结束语

本文讨论了一个基于规则方法的汉语到语义网络语言转换系统的实现，初步实现了知识获取的半自动化。整个转换过程对分词与依存句法分析有较高的依赖性，转换结果直接依赖于分词与依存句法分析的结果，由于缺少语料，目前还无法用实验数据对此方法进行衡量。为了能更好的表示自动生成 SNetL 关系式中两个概念节点之间的关系，还需要语义字典的支持，这是今后系统改进的方向。本文仅对句子内词语之间依存关系到 SNetL 关系式的转换进行了研究，然而依存关系存在于各层语法单位之间，包括单句、复句、句群之中；如何利用这些依存关系进一部分析单句、句子之间的语义关系也是非常值得探讨的问题。

致谢 本文使用的汉语依存关系树库、分词-词性标注模块、汉语依存句法分析模块均来自哈工大信息检索实验室，在此向提供免费共享资源的哈工大信息检索实验室的老师与同学表示由衷的感谢。

参考文献：

[1] 陆汝铃. 人工智能（上、下）[M]. 北京：科学出版社，1995.p.107-p.150

[2] George F. Luger. Artificial Intelligence: Structures and Strategies for Complex Problem Solving [M]. Pearson Education Limited, 2002 . p.197

[3] 王永庆. 人工智能原理与方法[M]. 西安：西安交通大学出版社，1998. p.307-p.310

[4] 陆汝铃. 专家系统开发环境[M]. 北京：科学出版社，1994. p.247-p.261