**文章编号：**1003-0077（2011）00-0000-00

# 基于概念层次网络（HNC）理论的专利汉英机器翻译系统*

朱筠，晋耀红

（北京师范大学中文信息处理研究所，北京 100875；

中国专利信息中心-北京师范大学机器翻译联合实验室， 北京 100875）

**摘要：**本文提出了一种基于概念层次网络（HNC）理论的纯规则汉英专利机器翻译系统。该系统将整个翻译过程分为分析、转换、生成三个模块：首先利用 HNC 理论提供的语义特征对源语言进行语义分析，得到源语言的语义分析树；其次在此基础上对语义分析树进行过渡转换，将其结构调整为目标语的合法表达形式；最后在转换树上进行目标语生成。系统目前主要针对专利文本进行翻译，采用的知识库和规则库均由人工建立，并利用约 6000 句语料对系统进行了各项指标的测试。

**关键词：**语义分析；过渡转换；语义特征；专利文本汉英机器翻译

**中图分类号：**TP391　　**文献标识码：**A

# A Chinese-English Patent Machine Translation System Based on the Theory of Hierarchical Network of Concepts

Yun Zhu, Yaohong Jin

(Institute of Chinese Information Processing, Beijing Normal University, Beijing 100875, China;
CPIC-BNU Joint Laboratory of Machine Translation, Beijing Normal University,
Beijing 100875, China)

**Abstract:** Compared with ordinary text, patent text often has more complex sentence structure and more ambiguity of multiple verbs. To deal with these problems, this paper presents a rule-based Chinese-English patent machine translation system based on the theory of Hierarchical Network of Concepts (the HNC theory). In this system, the whole procedure is divided into three main parts, the semantic analysis of the source language, the transitional transformation from the source language to the target language and the generation of the target language. The knowledge base and the rule set are obtained from manually analyzing the semantic features of a training set which contains more than 6 000 Chinese patent sentences, and a specific method of evaluation is provided during the experiment.

**Key words:** semantic analysis；transitional transformation；semantic features；patent machine translation

## 1 引言

为便于国际交流与合作，专利文献经常需要被翻译成多种语言。但近年来专利申请量不断增长，由此带来了日益繁重的翻译压力。为了应对这种压力，对专利文献的自动翻译已经成为机器翻译中一个重要的应用领域。自 2008 年起，NTCIR-7 将专利机器翻译列为研讨会评测项目之一。2011 年 3 月，北京师范大学中文信息处理研究所与中国专利信息中心合作成立了机器翻译联合实验室，共同开展面向专利语料的机器翻译的研究。

专利语言作为一种技术语言和法律语言的综合体，具有明显的特点：句子一般都较长，表达繁琐且严谨，格式相对固定。这些异于普通文本的特点给专利文本翻译带来了更多的困难。现有的专利机器翻译系统多从普通文本的翻译系统中发展而来，对专利翻译的难点没有找到很好的特殊处理办法，这使得翻译效果一般都低于普通文本的翻译。

为解决这一难题，本文提出了一种基于 HNC 理论的汉英机器翻译方法，利用 HNC 理论中

* **收稿日期：**2013-09-27　　**定稿日期：**2013-10-13

**基金项目：**国家高技术研究发展计划（863）（2012AA011104）；中央高校基本科研业务费专项资金

**作者简介：朱筠**（1983——），女，博士后，主要研究方向为中文信息处理；晋耀红（1973——），男，教授，主要研究方向为中文信息处理。

提供的语义特征，对专利文本进行分析、转换和生成。这一翻译方法主要包括三个模块：首先是对汉语的基本句群进行分析，得到源语言的 HNC 句法分析树；然后根据汉英表达习惯的差异，通过增减节点、调整节点顺序和修改节点上的形态特征，将源语言的句法树转换为合法的目标语句法树；最后在目标语生成树上生成翻译结果，输出为一个英文的完整句子。

由于英语与汉语表达习惯不同，汉语中的一个句末点号内常常包含多个具有完整句法结构的小句，但英语中一个简单句只能有一组主谓宾结构，否则就需要采用复杂句、复合句或不定式等形式表达。而专利的表述特点就是句子较长，包含多级句子结构。这些句子在翻译成英文时，需要进行句子结构的调整。因此专利汉英机器翻译应当以句末点号结尾的大句为基本处理单位。

本文将采用以下结构进行阐述：除引言和结论外，第二部分介绍相关研究工作，第三部分介绍系统中所使用的翻译策略，第四部分介绍翻译系统的各模块及调度流程，第五部分对系统进行测试并分析测试结果。

## 2 相关研究工作

目前，机器翻译主要采用基于规则的方法、基于统计的方法、基于实例的方法和基于混合策略的方法。基于规则的方法的优点在于规则可以直观、准确地描述语言现象，但规则的书写需要大量的时间和人力，另外规则的完备性很难得到保证，并且规则之间往往存在矛盾。基于统计的翻译方法本质上是利用噪声信道模型将翻译过程视为解码过程，关键是要选取最适合的语言概率模型和翻译概率模型，并根据语料对这些模型中的参数进行估计。这种方法的优点在于能够快速得到结果。但是由于数据稀疏等原因，对一些特殊的语言现象缺乏处理能力。统计方法在专利翻译中应用较广，根据 NTCIR-7 提供的数据，在那次研讨会中共有 15 支参赛队参与了专利翻译这个项目的评测，其中有 12 支队伍提交的系统都是采用了统计的翻译方法。现阶段，为了弥补这两种方法各自的缺点和不足，将两种翻译方法融合在一起已经成为机器翻译研究的一种新趋势。

在源语言分析阶段，谓语动词的识别对汉语句子的切分起到了很关键的作用。关于谓语动词的识别方法，北京大学的穗志方博士和俞士汶教授提出了基于双语语料库（穗志方，1998a）和基于决策树模型（穗志方，1998b）的两种汉语谓语中心词识别方法。有研究者提出了一个规则和特征学习相结合的谓语识别模型（龚小谨，2003）。有研究者利用主语和谓语之间的句法关系来识别谓语中心词（李国臣，2005）。

## 3 基于 HNC 理论的翻译策略

HNC 理论定义了一系列语义特征，这些语义特征不仅仅为本文的研究提供了理论支持，还为规则的书写提供了一组复杂特征集。本文中所使用的语义特征来自 HNC 理论中的 1v 准则，主要利用 1 类概念和 v 类概念对分析的各个层次进行切分。这里所指的“1 类概念”是广义的逻辑概念，包括 HNC 概念层次网络中的语言逻辑概念、语习逻辑概念、综合概念、基本概念等；“v 类概念”是动态概念。

本文对源语言的分析和转换都是利用规则方法进行实现的。规则系统受到质疑的一个原因在于，若规则描述过于简单，则规则产生的结果或者相互矛盾，或者不足以分析句子。若想完全依赖规则准确地给出分析结果，就需要每一条规则能够描写复杂的语言现象，这使得规则的概括性差，书写需要大量人工，不具有可行性。

为解决这一矛盾，我们引入边界感知原则，作为规则之上的规则进行调度。在此原则指导下，首先需要对规则进行层次分类，每一类规则只在固定分析层次中调用，且每一条规则只关注对邻近语串中语言现象的分析，不需要兼顾对整体形势的判断，而是通过调度来解决规则的兼容性问题。解决的策略有两条：首先避免规则的贪婪匹配，使规则调用具有层次性，

并在每一个层次上依据激活信息调用相应规则；其次，调度会根据不同处理阶段的句类特征对规则生成的结果进行选择合成。这样我们既减少了需要匹配的规则，也减少了不同规则所产生的矛盾对最终分析的影响，以此加强了对规则调用的控制，做到“模糊规则，精确调度”。

系统将句末点号结尾的大句分为三个层次进行分析：大句、小句和语块内部。构成大句的是小句和小句断句标记，其中每个小句具有独立的全局特征语块（谓语）；构成小句的是广义对象语块（主语和宾语）、全局特征语块（谓语）、辅语块（简称辅块，对应理论语言学中的状语）和 L 类语块。其中语块内部又可以根据构成方式不同，分为并联结构、串联结构和句蜕等。HNC 理论定义句蜕是由一个句子蜕化的语块或语块的一部分，如“参加这次会议的代表”是由“代表参加这次会议”这一小句蜕化而来。

### 3．1 语义特征

本系统中主要用到以下概念类别的语义特征：

- lb 类概念

lb类概念属于逻辑概念，主要对应于理论语言学中的连词，如并且、因为……所以……等，是用来句内断句和判断小句间关系的主要依据。

- l0 类概念

l0类概念是一类逻辑概念，它的出现常常改变谓语的施事和受事在小句中的位置。l0分为引导受事和施事，如“将”是引导受事的l0，它的出现常会将受事提前至谓语之前；“被”是引导施事的l0，它的出现常会将受事提至句首，将施事移至受事之后、谓语之前。

- l1 和 l1h 类概念

l1类概念标识了辅块的前边界，l1h标识了辅块的后边界。有些辅块两个边界都齐全，而有些辅块只有前后边界的其中之一。

- l4 类概念

l4类概念主要用于语块内部构成的分析，主要包括并联结构的标记“和、或”等连词，及串联结构和句蜕的标记“的”。

- 动态概念

除了以上l类概念，动态概念v也具有以下可以利用的语义特征。

HNC理论将动态概念分为广义作用和广义效应两大类，由广义作用类动态概念构成全局特征语块的小句属于广义作用句；由广义效应类动态概念构成全局特征语块的小句属于广义效应句。

根据动态概念的不同，构成小句的主语块数量（广义对象语块+全局特征语块的总数）可分别为2块、3块和4块，对应动态概念的语块特征分别为2、3、4。

另外，汉语中有些动态概念具有块扩属性，即该动态概念做全局特征语块时，可以紧邻一个宾语从句。

以上概念及语义特征是本文所述系统对源语言进行分析转换时的重要依据。除了这些语义特征本身，这些概念在句中还有层次的区分。例如，

例句 1 数据输出装置（7）根据控制器的命令，将由数据排序装置所排序的数据中的有效数据输出到装置外部。

例句 1 中有两个 L0 语块，一个是“将”，一个是“由”。L0 语块“将”参与小句的切分，切分结果为广义对象语块 GBK“数据输出装置（7）”+辅语块 ABK“根据控制器的命令”+L0 语块“将”+广义对象语块 GBK“由数据排序装置所排序的数据中的有效数据”+全局特征语

块 EG“输出到”+广义对象语块 GBK“装置外部”。而另一个 L0 语块“由”，是由“数据中的有效数据由数据排序装置所排序”蜕化而成的句蜕“由数据排序装置所排序的数据中的有效数据”内部的 L0 语块。因此，需要将 10 概念区分为小句层次和语块内部层次。

这一层次区分同样适用于其他 1 类概念和 v 类概念。对于 1 类概念，我们通过 LEVEL 属性值来判定，其值为 1 的是切分小句结构的构件，其值为 2 的为切分语块内部结构的构件；v 类概念的层次通过动态概念所具有的权值来进行区分。

### 3．2 源语言分析

在源语言分析阶段需要从三个层次进行处理：在句间关系层面，需要将以句号结尾的大句切分为若干小句，且判断这些小句间的语义关系。

在小句层面，需要正确切分出小句结构，即小句所包含的辅语块、广义对象语块、全局特征语块及引导这些语块的 L 类语块。

最后是对小句中拆分出来的这些语块进行语块内部构成形式的判断和展开，这一过程需要反复递归进行，直至句中所有语块都被拆分到单独一个词语形成的叶子节点，这类节点是句法树的终结符。

我们对规则的调用和结果处理主要遵循边界感知原则，根据处理阶段和对语串构成形式的预判调用相应规则，判断各类语块的边界、属性及层次信息，然后对这些边界信息进行选择、合成和挂树的操作。语块的边界判断主要依靠各类逻辑概念 l 和动态概念 v。由于我们对小句的定义是可以形成独立句法树，即具有独立的全局特征语块的语串。因此对小句的切分需要依赖小句内部全局特征语块选择的情况。所以，我们的处理顺序是小句分析与句间关系分析交叉进行。在这两个层次处理完成后，再逐层对语块内部的构成进行分析。

- 小句切分及句间关系分析

在小句切分及句间关系分析阶段，系统依次处理可以断句的逗号、分号、顿号、句号和 lb 类概念，在可以断句的位置生成预判的断句标记。由于小句的断句标准是包含独立的谓语，因此需要通过判断预判断句标记之前的语串是否包含全局特征语块，来决定是否在这个预判 SST 的位置生成最终断句标记 SST。在小句切分及句间关系分析完成时，整个大句被切分为若干小句与小句切分标记 SST 的集合。

$$CS=\sum(SS+SST)$$

- 小句分析

在小句分析阶段，主要通过规则生成语块边界并判定语块层次，然后调度程序对语块边界进行合成排序，利用属于小句层次的动态概念和 l 类概念（包括 l0，l1 和 l1h），将小句切分为辅块（ABK）、广义对象语块（GBK）、全局特征语块（EG）和各类起引导作用的 L 类语块。在小句分析完成后，小句可被切分为以下格式：

SS = ABK [alternative] + GBK [alternative] + EG + GBK + GBK [alternative]

SS = ABK [alternative] + GBK [alternative] + l0 + GBK + EG + GBK [alternative]

其中 ABK = l1[alternative] + 语串 + l1h [alternative]，l1 和 l1h 必选其一。

- 语块内部构成分析

语块内部构成分析规则是在小句句法树生成之后，对树上切分出来的广义对象语块 GBK 进行下一层次的分析切分处理。GBK/BK 语块内部可由串联结构、并联结构和句蜕构

成。有的语块只包含一种结构，且构成方式较为简单；还有很多语块具有复杂的构成形式，可能具有多种结构的嵌套和交叉。因此需要结合待分析语块的内部、外部的语义特征，对语块构成形式做出预判。然后根据预判结果，制定规则调度策略。语块分析的调度及规则是需要不断递归调用的，直到语块被拆解为不能再切分的终结符为止。经过分析，语块可被切分为：

串联结构

串联语块 1+串联标记“的”+……+串联标记“的”+串联语块 n；

并联结构

并联语块 1+并联标记“、/和/或”+……+并联标记“、/和/或”+并联语块 n；

句蜕

句蜕修饰成分+句蜕标记“的”+句蜕中心成分。

### 3．3 过渡转换

由于汉语和英语在表达习惯上存在一些差别。因此在对源语言分析完成后，需要经过汉英过渡转换处理，以便得到合乎英文表达规范的句子。转换处理主要通过添加/删减树节点、调序和为树节点赋属性值这四种操作，将源语言分析树转换为合法的目标语的句法树。

HNC 理论中的过渡转换处理共分为六个环节，包括“两转换、两变换、两调整”——句类转换环节、句式转换环节、主辅语块变换环节、语块构成变换、辅块排序调整和小句排序调整。若在整棵树上同时进行六项转换操作，则太过复杂，不易操作。因此需要将这些表达差异分层次分步骤进行转换。如果只关注某一层次某一环节的转换操作时，就只需要关注某一层次的树节点，而不需要关注整个树结构，也不需要关注某个节点的内部结构。利用这一特点，也为了降低处理的难度，可利用树结构所具有的层次性，分层进行转换。这样，在每个处理的阶段，只需要关注处于同一层次的兄弟节点。将这些节点按照兄弟关系排列为一个有序序列，在这个有序序列上进行转换操作。

有时，处理所需的信息需要利用树节点之间的父子关系进行传递，这种传递是双向的。对于转换时所需的子节点信息，在分析树生成阶段会被抽取到相应的父节点上。而当转换结果需要改变子节点的特征时，只需赋值给父节点，然后通过父子节点之间的信息传递实现最终的转换。

因此，过渡转换模块同样包含三个层次，即句间关系转换、小句结构转换以及语块内部度转换。每一次转换操作是在拥有共同父亲节点的有序序列上展开的。为不改变原始分析结果，需要对一个树节点所在层次处理完后，在对该节点内部进行转换。因此，转换规则需要按照由高层树节点到低层树节点，由大范围到小范围这样的顺序展开。

- 句间关系转换

句间关系转换主要是根据小句间的关系，将大句中的各个小句组合成合法的英文长句形式：有些小句需要添加关联词引导，有些小句需要转换为从句或非限定性动词短语的形式。对于需要转换为非限定性动词短语的小句，需要将谓语动词需采用的形态（过去分词或现在分词）属性赋值到小句节点上。

- 小句结构转换

小句结构转换主要是确定小句各构件语块的顺序及全局特征语块中动态概念应采用的合法形态，包括辅块位置的调整，格式转换（广义作用句且包含 l0 语块的小句顺序调整及谓语语态变换），样式转换（广义效应句及不包含 l0 语块的广义作用句的小句顺序调整及谓语语态变换），全局特征语块中动态概念的时态、体态、语态以及非谓语形式等形态特征的

确定。

- 语块内部结构转换

语块内部的结构转换主要是在串联结构和句蜕中，将原位于语块块尾的中心成分调整至块首，而将修饰部分调整至中心成分之后。

### 3．4 目标语生成

目标语的生成是在经过过渡转换的句法树上展开。在这部分工作中，主要需要根据词语所处在语块的位置以及临近词的语义特征，确定兼类词应采用的概念类别；然后根据最终确定的概念类别在汉英对照翻译词典中进行选词；最终根据过渡转换中赋予的形态特征对选定词汇进行变形输出。

## 4 系统各模块及调度流程

北京师范大学中文信息处理研究所与中国国家专利信息中心合作，成立了机器翻译联合实验室，集中力量对专利文献的汉英机器翻译系统进行攻关。依照上文所介绍的处理方法，搭建基于 HNC 理论的面向专利文本的汉英机器翻译系统。系统主要包含四个主要模块：分词模块、源语言分析模块、过渡转换模块和目标语生成模块，如下图所示：

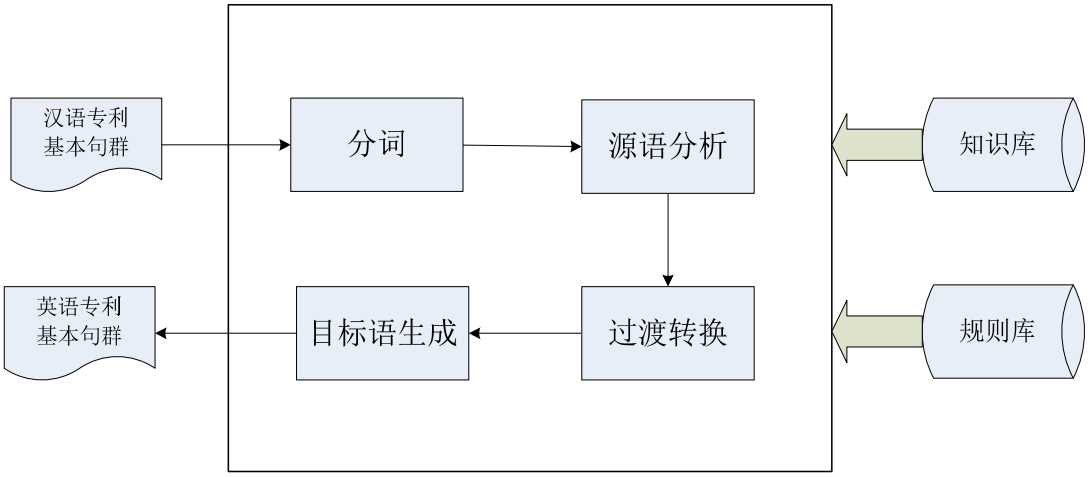


图 1: 专利汉英机器翻译系统总流程图

（1）分词模块主要是依据知识库中给出的词形对汉语专利中以句末点号结尾的大句进行分词。

（2）源语分析模块主要依据知识库提供的各项知识与分析规则库对源语言进行分析，得到源语言大句的句法分析树。

（3）过渡转换模块主要依据句法分析树以及转换规则库，通过对句法分析树进行调序、增加节点、删除节点和更改节点属性这四项操作，将汉语句法分析树转换为符合英文表达习惯的目标语句法树。

（4）目标语生成模块是根据过渡转换后的目标语句法树的节点次序和属性，选择合适的对译词并进行变形，生成英文专利的翻译结果。

### 4．1 分析模块的调度流程

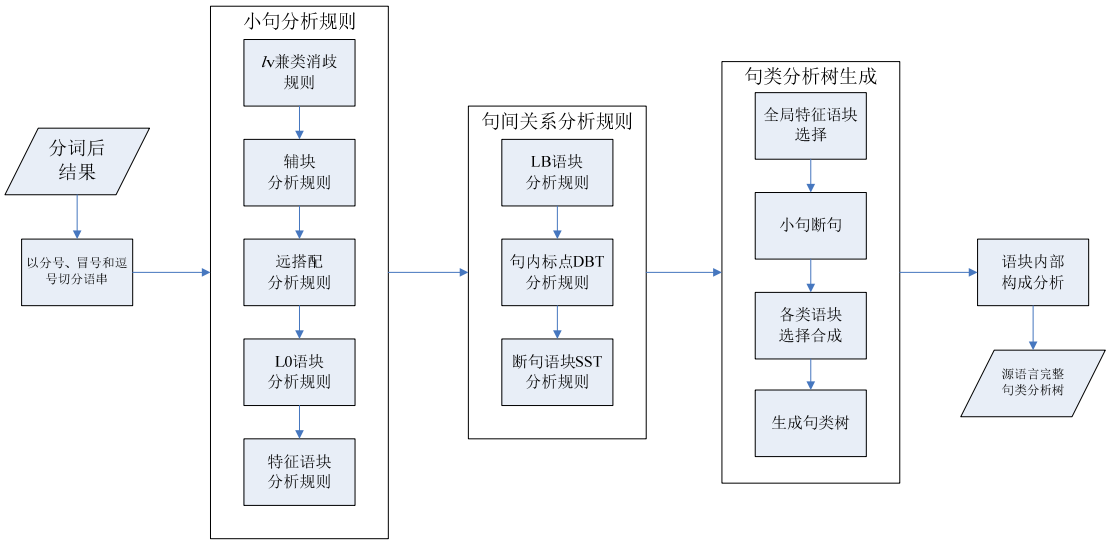


图 2: 源语言分析模块流程图

分析模块按照以下顺序进行调度：

（1）利用逗号、分号和冒号将分词后的基本句群切分为若干语串。

（2）调用小句内部的分析规则，识别各类 L 类语块和特征语块的边界及层次。

（3）调用句间关系的分析规则。

（4）生成句间和小句层的句法分析树：先合成规则标记的语块边界，从标记的结果中选择全局特征语块，根据特征语块判断小句是否断句，然后选择与全局特征语块搭配的合法小句构件语块进行合成挂树。

（5）在句法树上，逐层扫描广义对象语块 GBK 进行语块内部构成的分析。

至此完成源语言分析，生成完整的源语言句法分析树。

## 4．2 转换模块的调度流程

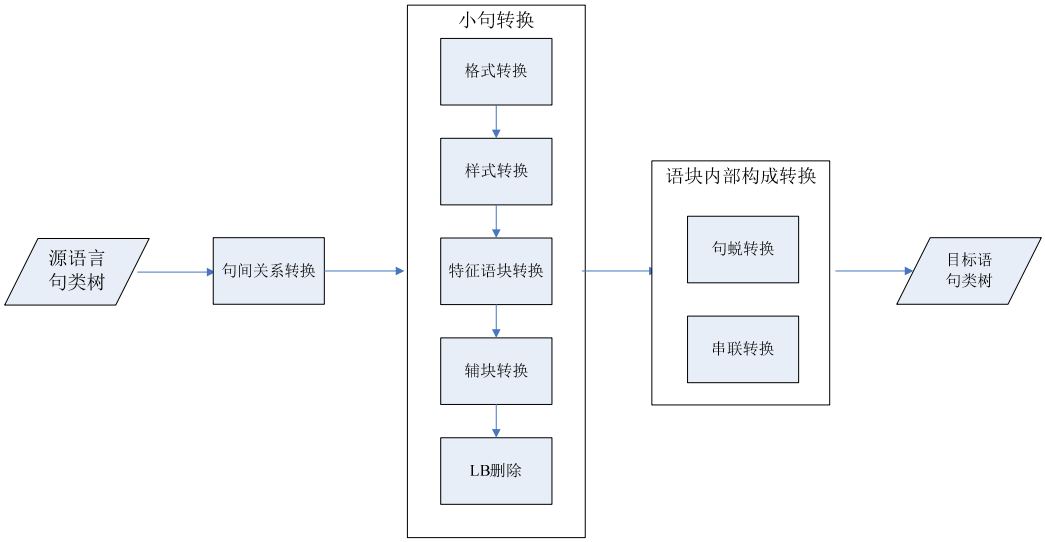


图 3: 过渡转换模块流程图

转换模块按照以下顺序进行调度：

（1）将源语言生成树送入句间关系转换环节，实现过渡转换中的部分句式转换和小句句序的调整。

（2）将句法分析树上的所有小句交由小句转换环节进行转换。
（3）将句法分析树上的所有广义对象语块语块 GBK 交由语块内部构成转换环节进行转换。在这一环节中，每个待处理语块只需调用一种类型的转换规则。

至此，完成从源语言句法分析树向目标语句法分析树的转换。

### 4．3 生成模块的调度流程


图 4: 目标语生成模块流程图

在过渡转换模块完成后，对目标语句法分析树上的每一个叶子节点选择合适的概念类别，然后将汉语词性和概念类别作为查找条件，在知识库中找到对译词语，根据节点上标注的形态特征，对对译词语进行变形，将所有处理过的词语按句法树的顺序整理成英文的翻译结果。

至此，系统完成了将一个汉语专利基本句群翻译为英语的过程。

## 5 测试及结果分析

实验所用专利汉英机器翻译系统的知识库由人工构建，共包含 53291 个词语，覆盖了近 500 篇真实专利文献中出现的所有词语，知识库包括词形及词语所具有的概念知识和句类知识。

规则库的搭建首先从中国专利信息中心提供的 277 句典型双语对照专利语料入手，依据语料中的典型语言现象编写相应规则。之后随着调试工作的深入，对规则库进行不断的修正和补充。现在，分析规则库共包含 31 个文件，1183 条规则；转换规则库共包含 10 个文件，439 条规则。

由于本系统所采用的处理方法主要注重对源语言的分析和转换，因此对于系统的测试，在调试阶段未采用常用的 BLEU 值作为测量指标，而是将处理过程中所关注的主要语言点作为测试点，计算各个测试点的准确率和召回率。

由于小句中特征语块的识别正确与否对小句结构的分析起到至关重要的作用，因此首先针对 277 句典型语料的全局特征语块辨识效果进行了封闭测试，并将测试结果与著名的线上机器翻译系统 Google 给出的结果进行了对比。由于 Google 在线翻译直接给出了翻译结果，因此表中数据是根据翻译结果中的谓语动词进行反推计数的。

| | **总数** | **总识别数** | **正确识别数** | **正确率 P(%)** | **召回率 R(%)** | **F(%)** |
|---|---|---|---|---|---|---|
| HNC 系统 | 363 | 354 | 333 | 94.1 | 91.7 | 92.9 |
| GOOGLE | 363 | 288 | 217 | 75.3 | 59.8 | 66.7 |

表 2: 全局特征语块识别的封闭测试结果

在 277 个句子中，共包含 363 个全局特征语块，利用本文所述方法识别的全局特征语块的数量是 354 个，其中识别正确的数量是 333 个；Google 识别的全局特征语块总数为 288 个，识别正确的数量是 217 个。由这些数据计算可知，我们的系统对全局特征语块的识别正

确率为 94.1%，召回率为 91.7%，F 值为 92.9%。在全局特征语块识别这个语言点上，我们的系统在封闭测试中取得了较好的成绩。

接下来我们将测试点进行扩充，在分析阶段选取小句切分和全局语块辨识两个测试点，在转换阶段选取格式转换和辅块转换两个测试点。仍然先对 277 个基本句群进行了封闭测试，测试结果如下：

| 分析阶段 | | 转换阶段 | |
|---|---|---|---|
| 小句切分正确率 | 全局特征语块辨识正确率 | 格式转换正确率 | 辅块转换正确率 |
| 92% | 94% | 80% | 90% |

表 2：针对四项测试点的封闭测试结果

之后我们选取了知识库已覆盖的 30 篇完整的专利文本，共包含 6308 个大句，测试点不变，对系统进行开放测试。下表是开放测试 30 篇语料测试结果中的最好结果、最差结果和平均结果。

| | 句子总数 | 分析阶段 | | 转换阶段 | |
|---|---|---|---|---|---|
| | | 小句切分正确率（%） | 全局特征语块辨识正确率（%） | 格式转换正确率（%） | 辅块转换正确率（%） |
| 最好结果 | 267 | 96 | 91 | 94 | 96 |
| 最差结果 | 252 | 68 | 46 | 43 | 80 |
| 平均结果 | 6308 | 83 | 75 | 74 | 85 |

表 3：针对四项测试点的开放测试结果

由结果可以看出，本文所述的处理方法对各项测试点的平均正确率基本达到了 75%的水平。造成分析转换中的问题主要有以下几个原因：

1. 知识库有待修正。由于本系统是纯规则系统，处理时所需的很多知识都依赖知识库和规则库的准确性。若知识库提供了错误或者不全面的信息，处理时就会产生错误。

2. 规则库的修正与补充。同知识库一样，规则库也需要通过调试语料进一步修正现有规则并补充新的规则，以便提高处理的正确率和召回率。即使如此，我们依然可以看出，依据 1622 条规则，已经可以达到现在的处理水平，这可以说明现有规则具有较好的覆盖性。

3. 处理策略的细化和调整。除了对知识库和规则库的依赖之外，有些错误处理结果需要通过进一步细化调整规则来实现。如并联结构的识别效果还不够理想，有些分析错误的情况不能通过规则来解决，此时就需要增加新的识别方法，调整现有策略，以改善效果。

4. 程序中的错误。有些处理错误的情况与程序有关，需要通过调试不断消除程序中的错误，减少程序对处理结果的负影响。

## 6 结论

本文针对中文专利语料的特点，设计了一种利用语义特征构造规则，以此对汉语专利文献进行翻译的方法，并利用该方法构建机器翻译系统系统。通过针对一些特殊的语言点对该系统进行封闭测试及开放测试，正确率基本能够超过 75%，效果较为理想。

在未来的工作中，笔者还需要在大规模的测试语料上对系统进行不断测试，根据测试结果继续完善系统各个模块的调度策略，提高系统的性能。同时，要加强系统对其他语义特征

的分析和利用。

目前系统还在对各类语言现象处理效果的调试中，着重调试源语言的分析和语序的调整。但生成模块还未进行仔细调试，对翻译结果的后处理也还未开展，最终结果还不够理想。因此并未对系统进行 BLEU 值的评测。在进一步完善以上模块后，将选取机器翻译中常用的评测方法，对系统进行评测。

## 参考文献


[1] Fujii, M Utiyama, M yamamoto, et al. Overview of the Patent Translation Task at the NTCIR-7 Workshap[C]//Proceedings of NTCIR-7 Workshop Meeting, Japan. 2008.

[2] X. Dai, C Yin, J. Chen, G. Zheng. Machine Translation: Past, Present, Future[J]. Computer Science, 2004,Vol. 31, No. 11, pp.176–179, pp.184

[3] 黄曾阳. HNC(概念层次网络)理论[M]. 北京：清华大学出版社. 1998.

[4] 苗传江. HNC（概念层次网络）理论导论[M]. 北京：清华大学出版社. 2005.

[5] 晋耀红. 基于 HNC 理论的句类分析系统的设计与实现. 见: 黄曾阳. HNC(概念层次网络)理论. 北京：清华大学出版社. 1998.

[6] 穗志方，俞士汶. 面向 EBMT 的汉语单句谓语中心词识别研究[J]. 中文信息学报. 1998a. Vol.12 No.4 P39-46

[7] 穗志方，俞士汶. 汉语单句谓语中心词识别知识的获取及应用[J]. 北京大学学报(自然科学版). 1998b. 第 34 卷 第 2-3 期 P221-229

[8] 龚小谨，罗振声，骆卫华. 汉语句子谓语中心词的自动识别[J]. 中文信息学报 2003. Vol.17 No.2 P7-13

[9] 李国臣，孟静. 利用主语和谓语的句法关系识别谓语中心词[J]. 中文信息学报 2005. Vol.19 No.1 P1-7

[10] 韦向峰，熊亮，张全. 汉语语句核心动词的自动获取研究[J]. 计算机工程与应用 2007. 43( 10) P179-182



作者联系方式：朱筠 北京市海淀区新街口外大街 19 号 北京师范大学中文信息处理研究所 100875 电话：13810103477 电子邮箱：zhuyun@bnu.edu.cn